

HARVARD UNIVERSITY  
Graduate School of Arts and Sciences

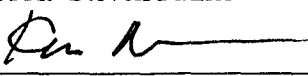


DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the  
Department of Psychology  
have examined a dissertation entitled  
**"The Causes of Individual Differences"**  
presented by **James J. Lee**

candidate for the degree of Doctor of Philosophy and hereby  
certify that it is worthy of acceptance.

Signature   
Typed name: Prof. Steven Pinker

Signature   
Typed name: Prof. Ken Nakayama

Signature   
Typed name: Prof. George Church

Signature   
Typed name: Prof. Thomas Bouchard

Date: September 6, 2011



The Causes of Individual Differences

A dissertation presented by

James J. Lee

to

The Department of Psychology

in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy  
in the subject of  
Psychology

Harvard University  
Cambridge, MA

September 2011

UMI Number: 3491882

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3491882

Copyright 2012 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

© 2011 — James J. Lee  
All rights reserved.

## The Causes of Individual Differences

### General Abstract

The aim of personality psychology is to explain the causes and consequences of variation in behavioral traits. The three papers collected in this dissertation attack this problem using a variety of empirical and theoretical approaches. (1) The first paper presents the results of a genome-wide association study of over 100 human phenotypes in a sample of 401 participants. The study failed to find any genetic variants significantly associated with the personality traits. Research on twins and other kinships have shown that these traits are highly heritable, and thus the study supports the view that their heritabilities are attributable to many loci of small effect. Drawing on Fisher's geometric model of adaptation, I offer the hypothesis that the different evolutionary trajectories of the traits examined in the study account for their disparate genetic architectures. (2) The second paper reports several studies focusing on the personality trait of general intelligence and its negative correlation with reaction time in elementary laboratory tasks. The studies found that the correlation is attributable to the time taken by a serial decision-making stage; the parallel perceptual and motor stages surrounding this serial stage do not contribute to the correlation. If this association between intelligence and speed of serial processing reflects a causal relationship, then it paves the way for a mechanistic understanding of ability variation at the neural and cognitive levels. (3) Personality research has long been dogged by controversies over the extent to which causal inferences can be drawn from observational data. In recent years the computer scien-

tist Judea Pearl has used a graphical approach to extend the innovations in causal inference developed by the population geneticists Ronald Fisher and Sewall Wright. Besides shedding much light on the philosophical notion of causality itself, this graphical theory now contains many powerful concepts of relevance to the controversies just mentioned. The third paper applies Pearl's theory to areas of personality research where questions of causation arise. In one part of the paper, I reanalyze a dataset bearing on the question of whether intelligence is cause of social liberalism.

# Contents

## **1 A Genome-Wide Association Study of 100+ Physical and Behavioral Traits Finds**

<b>Few Loci of Large Effect</b>	<b>12</b>
1.1 Introduction . . . . .	13
1.2 Results . . . . .	14
1.3 Discussion . . . . .	21
1.4 Materials and Methods . . . . .	30
1.4.1 Ethics Statement . . . . .	30
1.4.2 Participants . . . . .	31
1.4.3 Procedure . . . . .	44

## **2 $g$ and the Psychological Refractory Period: Individual Differences in the Mind's Bottleneck**

<b>2.1 Introduction . . . . .</b>	<b>53</b>
<b>2.2 Results . . . . .</b>	<b>58</b>
2.2.1 Analysis of Single-Task RT Means and Variances . . . . .	60
2.2.2 Diffusion Decomposition of Single-Task RT . . . . .	66
2.2.3 Analysis of Dual-Task Means, Variances, and Correlations . . . . .	70
2.3 Discussion . . . . .	85
2.4 Materials and Methods . . . . .	91
2.4.1 Participants . . . . .	91
2.4.2 Procedure . . . . .	92
2.4.3 Stimuli . . . . .	94



2.4.4	Data Analysis . . . . .	94
<b>3</b>	<b>Correlation and Causation in the Study of Personality</b>	<b>99</b>
3.1	A Unifying Theory of Causality . . . . .	107
3.1.1	The Interpretation of a Causal DAG . . . . .	107
3.1.2	The Value of Randomization . . . . .	121
3.2	The Nature of Psychometric Common Factors . . . . .	123
3.3	The DAG As a Source of Severe Empirical Tests in Structural Equation Modeling . . . . .	129
3.4	Concepts of Genetics . . . . .	140
3.4.1	Foundations of Heritability . . . . .	140
3.4.2	Ancestral Confounding . . . . .	140
3.4.3	Causal Inference in Gene-Trait Association Studies . . . . .	149
3.5	Conclusion . . . . .	155

*I would rather discover one causal law than be King of Persia.*

—DEMOCRITUS

*FOR MINJI AND MIRANDA*

## General Introduction

Personality psychology is concerned with variation in behavioral traits (Ashton, 2007). Notwithstanding the great achievements of the discipline in the last century, a number of fundamental barriers have hampered its further progress. These barriers can be grouped under three broad headings:

1. technological barriers to the direct observation of theoretical entities;
2. disciplinary barriers to integration with other relevant branches of psychology; and
3. conceptual barriers to drawing casual inferences from observational data.

Each of the three papers collected in this dissertation addresses one of these problematic areas. Causal inference, the last of these problems, is actually closely related to the previous two; it provides a perspective from which all three problems appear to be facets of the same problem.

I will discuss these problems in their enumerated order.

**The Direct Observation of Theoretical Entities (Genetic Variation)** In the past many of the variables hypothesized to be responsible for causing individual differences could not be “directly observed.” For example, when reading Fisher’s (1930) classic *The Genetical Theory of Natural Selection*, we should remember that at the time a “gene” was an abstraction whose existence was deduced from certain regularities in the inheritance of phenotypes. That is, no one knew what kind of physical entity a gene might be. Such a grounding had to await the observations using X-ray crystallography that culminated in the discovery of

DNA's double-helical structure (Watson & Crick, 1953). In the history of science, there are many similar examples of improvements in measurement technology enabling enormous empirical and theoretical advances (Gribbin, 2002).

We are arguably experiencing another such wave of technological improvements: the rapidly declining cost of measuring genetic variation at the DNA level. This decline has already allowed geneticists studying diseases and anthropometric traits to carry out *genome-wide association studies* (GWAS) in which *single-nucleotide polymorphisms* (SNPs) scattered throughout the genome are tested for association with a phenotype of interest (Wellcome Trust Case Control Consortium, 2007; McCarthy et al., 2008). A recent spectacular example of this approach is the discovery of over 180 genomic regions containing a variant affecting height in a sample of 180,000 individuals (Lango Allen et al., 2010). Moreover, a comparison of alleles associated with increased height found that they are more common in Northern Europe than in Southern Europe (Turchin, 2011). Interestingly, the magnitude of the difference in allele frequencies is correlated with the effect size of the height association, providing evidence that the divergences have been driven by natural selection specifically for body size. Population-genetic considerations suggest that this selection must have occurred within the last 10,000 years, prompting a number of intriguing hypotheses regarding the adaptive pressures driving the divergence (Bellwood, 2005; Anthony, 2007; Cochran & Harpending, 2009). This astounding series of findings provides a strong motivation for applying genome-wide association studies to personality variation. Knowledge of the genetic architecture underlying cognitive abilities, cooperation, and other key facets of human behavior may shed light not only on proximate biological mechanisms but also the ultimate evolutionary forces that have shaped the commonality and diversity of humankind. The first

paper in this dissertation takes a small step toward this goal by reporting a GWAS of several human phenotypes, including many traits of interest to behavioral scientists.

The results of the first paper are negative in the sense that no significant associations with personality traits were found. I claim that these results are owed to inadequate sample size to discover the “typical” causal variant under the influence of natural selection. But these results might lead a skeptical outsider to ask: What reason is there to believe that the genetic variants hypothesized to affect personality do in fact exist? How can we be sure that the increasing the sample sizes of personality genome-wide association studies will not prove to be a wild goose chase? Personality psychologists believe that the traits that they study are heritable because of certain patterns in the correlations between relatives (Fisher, 1918; Falconer & Mackay, 1996; Lynch & Walsh, 1998; Visscher et al., 2008). Genetic theory provides strong constraints on these correlations, and from their numerical values we can estimate the proportion of the trait variance caused by genetic differences. Applications of these methods to personality traits have led to reports of substantial heritability (Bouchard & Loehlin, 2001; Bouchard & McGue, 2003), which might seem to justify the expansion of genome-wide association studies to these traits. It is a rather curious sociological fact, however, that many geneticists do not believe these heritability estimates. One prominent human geneticist has said, “I won’t believe that there are genes for being smarter until you point me to the actual genes.” At meetings and presentations where the heritability of personality is raised, human geneticists and other biologists continue to assail these findings, claiming either that the data are fraudulent or that nothing worthwhile can be concluded from them.<sup>1</sup>

---

<sup>1</sup>This apparently widespread disbelief is curious because the exact same methods—studies of twins, parents-children, adoptees, and other kinships—are routinely relied upon by the human genetics community to provide estimates of “missing heritability” benchmarking the progress of genome-wide association studies in

In summary, for reasons that are not entirely clear, the available evidence from the correlations between relatives has failed to convince many scientists that genetic differences are an important cause of personality differences. It is thus apparent that genome-wide association studies of behavioral traits may be necessary, not only to advance personality psychology to new frontiers, but also to solidify its past achievements. However, without addressing the important issue of causal inference in genome-wide association studies,<sup>2</sup> there is a danger that skepticism toward gene-trait causation will simply transfer from traditional biometrical studies to molecular studies. Biometrical studies of twins and other kinships rely on an “indirect” chain of inferences from the correlations between relatives to their causal sources that was worked out by Fisher (1918) well before the discovery of heredity’s molecular substrate, and it is perhaps understandable that modern geneticists feel uneasy about causal inferences within a framework where genes need be nothing more than invisible and weightless theoretical entities. Turkheimer (2008) has argued, however, that the seeming straightforwardness of examining correlations with “direct” measures of genetic variation should not excuse genome-wide association studies results from being regarded with similar disdain. His argument is essentially that the techniques employed in genome-wide association studies—multiple regression, principal components, within-family designs, null hypothesis testing, and so forth—have proved to be inadequate positive tools of causal inference in the social sciences and therefore, by analogy, their applications in genomics will prove to be just as worthless.

---

identifying loci associated with disease and anthropometric traits (Manolio et al , 2009)

<sup>2</sup>Here I am not referring to the problem of isolating the precise causal variant within a genomic region associated with a phenotype (Pomerantz et al., 2009; Stacey et al , 2010; Musunuru et al , 2010, Holm et al , 2011). I am referring to the logically prior problem of determining whether an associated region contains a causal variant at all

Turkheimer's warnings have been entirely disregarded by the human genetics community. Even those who criticize genome-wide association studies on other grounds agree that the great majority association signals reflect the causal effects of nearby markers (Goldstein, 2011b). But what is wrong with Turkheimer's argument? What justifies the confidence of complex trait geneticists? Unavoidably we are led to the more general problem of drawing causal inferences from observational data, a major topic of the third paper in this dissertation. In that paper the specific context of genome-wide association studies is discussed at some length.

**The Integration of Personality and Experimental Psychology** Given the bare fact that a personality trait is heritable, there is still a causal chasm between the precise genetic variants affecting the trait and measurements of the trait itself. The purpose of GWAS is obviously to narrow this chasm from the genetic side. We can also begin narrowing the chasm from the phenotypic side by seeking to explain high-level personality traits in terms of more basic neural properties or psychological constructs. Such a reductionistic endeavor seems to require a closer integration of personality psychology (the study of individual differences) and experimental psychology (the study of species-typical behavior), since it is the latter that is concerned with the mind as a causal system.

Unfortunately, throughout the twentieth century, a deep divide persisted between personality and experimental psychology. In his Presidential Address to the American Psychological Association, Cronbach (1957) decried the separation of "the two disciplines of scientific psychology" and called for their unification. Before the time of Cronbach's address, however, a shotgun marriage of the two psychologies would probably have proven

barren. Behaviorism, the reigning paradigm in experimental psychology through World War II, focused on the relations between stimuli (causes) and responses (effects) while neglecting any intervening mental structure. It is the postulation and validation of such a qualitative structure, however, that supports an interface with personality psychology. It is then natural to hypothesize that quantitative variation in certain elements of the structure gives rise to the individual differences that personality instruments record.

At the time of Cronbach's address, behaviorism had begun to give way within experimental psychology to cognitivism, which is committed to the existence of mental structure (Pylyshyn, 1984). In the last few decades, cognitive psychologists have discovered several striking facets of this structure that may serve as a foundation for the research program that Cronbach envisioned. The second paper in this dissertation begins building such a foundation by exploring the nature of the negative correlation between IQ scores and reaction time (RT) in elementary cognitive tasks. Specifically, I test the hypothesis that the IQ-RT correlation is due solely to a serial decision-making stage that maps the stimulus to the appropriate response (Pashler, 1998; Sigman & Dehaene, 2005, 2006). The corollary is that the parallel perceptual and motor stages surrounding this decision-making stage make no contribution to the IQ-RT correlation. I test this hypothesis using several distinct techniques developed by experimental psychologists following the cognitive revolution (Sternberg, 1969; Ratcliff, 1978; Pashler, 1994), suitably extending these techniques when necessary to deal with individual differences. Thus, the studies in the second paper depart from other recent attempts to use constructs of experimental psychology in intelligence research (e.g., Conway et al., 2007); unlike these other attempts, the studies make intimate use of experimental methods to test precise and telling predictions. Other approaches that simply take "working mem-



ory capacity,” “executive control of attention,” and the like, treating them indistinguishably from psychometric common factors in a purely correlational study, are arguably much less revealing.

The studies reported in the second paper take two basic approaches. The first approach is to subject RT to various experimental manipulations, each posited to affect a distinct stage, and determine whether those manipulations known to affect the serial decision-making stage also show a privileged relationship with IQ (Sternberg, 1969; Wagenmakers et al., 2007; Grasman et al., 2009). The second approach is more direct. By presenting two stimuli in each trial and varying the time between their onsets, one can infer from the manner in which the two responses interfere with each other whether the IQ-associated stage is a serial stage.

There are several reasons why isolating the IQ-RT correlation to a particular processing stage would advance the subbranch of personality psychology concerned with ability differences closer to Cronbach’s call for a unified discipline. First, the temporal position and time-sharing properties of the IQ-associated stage would suggest several fruitful avenues for integrating individual differences with theoretical accounts of problem solving, analogy making, “mental modeling,” and other multistep cognitive processes (Carpenter et al., 1990; Hofstadter & the Fluid Analogies Research Group, 1995; Johnson-Laird, 2006). Second, it has already been shown that the serial decision-making stage contains a noisy accumulation of evidence picking out one of a few alternatives (Sigman & Dehaene, 2005, 2006; Ratcliff et al., 2008), which is suggestive of the neural mechanism by which this stage computes the appropriate discrete response to an analog input (Wong & Wang, 2006; Gold & Shadlen, 2007).

Tracing the explanatory chain forward from the genes, and backward from behavior through cognitive architecture and the brain, should eventually result in a consilient meeting. A problem that haunts this entire enterprise, however, is the nature of the connection between causality and observational data. Confirming the hypothesis of an exclusive association of IQ with a serial, stochastic decision-making RT stage offers the promise of explanatory progress only if this association reflects a *causal* effect of individual differences in this stage on complex thought processes in a variety of settings. But how might this leap from association to causation be made? Experimental manipulations can rule out certain stages as the source of the IQ-RT correlation, but they do not by themselves support the remaining stage as playing a causal role.

Some psychologists claim that the *only* study design supporting causal inferences is the random assignment of participants to different levels of the putative causal variable (Nisbett, 2009; Chabris & Simons, 2010).<sup>3</sup> This claim reflects a deep antipathy of experimental psychologists toward the goals and methods of personality psychology that Cronbach discussed at some length in his address. This skepticism toward whether personality psychology is indeed a science of causes and effects has even been abetted by some personality psychologists themselves. Under the influence of Edwardian scientists who thought that the notion of causality could not be formulated mathematically, these personality psychologists have denied that causation is what their non-experimental methods aspire to demonstrate (Burt, 1940; Lubinski & Dawis, 1995). These writers would have us think that one of Cronbach's two psychologies is a "science of causes" and the other a "science of correlations"; the rele-

---

<sup>3</sup>It is true that these authors wield this claim rather selectively, citing observational data whenever they happen to support their favored theories.

gation of personality psychology to the latter is then asserted to be a virtue!

My own claim is that if the extreme positions just summarized are correct, then a research program beginning with the association between IQ and a particular processing stage must quickly run into a dead end; it is difficult to imagine the circumstances under which the relevant neural variables could be identified and experimentally controlled. It is clear, however, that these extreme positions are untenable. First, they cannot account for the many examples from the physical sciences (the moon causing the tides) and epidemiology (tobacco causing lung cancer) where widely accepted causal inferences *have* been drawn from observational data. Second, they come dangerously close to *defining* causation as a difference following randomization rather than treating randomization as a tool for *revealing* causation.

The further integration of personality and experimental psychology would appear to benefit greatly, then, from a systematic and explicit framework that both defines causality without reference to randomization and explains the effectiveness of randomization as a tool. Such a framework is the subject of the third paper.

**Causality and Observational Data** Each of the first two papers concerns a non-manipulable hypothesized cause of individual differences. In both papers I present evidence that is *consistent* with the causal hypothesis (although in the first paper this evidence does not extend beyond physical traits such as eye color). How *strongly* is the causal hypothesis supported, however, by the presented evidence or evidence that might be gathered in future studies? The answer to this question depends on considerations only briefly sketched in the papers themselves. The third paper in this dissertation is an extended exposition of these considerations.

The notion of *association* is precisely captured by the concept of *conditional probability* in probability theory. Scientists are so used to working with conditional probability (expressed variously as a correlation coefficient, regression coefficient, and so forth) that it may come as a surprise that the notion of *causation* is not at all embraced by this concept.

The word *cause* is not in the vocabulary of standard probability theory. It is an embarrassing yet inescapable fact that probability theory, the official mathematical language of many empirical sciences, does not permit us to express sentences such as “Mud does not cause rain”; all we can say is that the two events are mutually correlated, or dependent—meaning that if we find one, we can expect to encounter the other. *Scientists seeking causal explanations for complex phenomena or rationales for policy decisions must therefore supplement the language of probability theory with a vocabulary for causality, one in which the symbolic representation for the causal relationship “Mud does not cause rain” is distinct from the symbolic representation for “Mud is independent of rain.”* Oddly, such distinctions have yet to be incorporated into standard scientific analysis. (Pearl, 2009, p. 134, emphasis added)

In this passage the computer scientist Judea Pearl rightly points out that causality has remained a merely informal, non-mathematical concept despite its vital scientific importance. As evidenced by the discussion of the first two papers, such informality can only be a stumbling block in fields (such as personality psychology) where attempts at causal inference have attracted controversy. Building on the foundations laid by the population geneticists Ronald Fisher and Sewall Wright, the computer scientist Judea Pearl has achieved the remarkable feat of formalizing causal reasoning in a manner that is both deep and accessible.

The third paper in this dissertation shows how Pearl’s theory provides tools to clarify the problems of causal inference arising in personality psychology. As just one example of this theory’s power, consider the problem of what variables to statistically control in order to obtain an unbiased estimate of a causal effect using multiple regression (or some closely re-

lated method). The typical student's methodological training will include advice to the effect that one should control for all measured variables that are correlated with both the putative cause and effect. Although this advice was criticized by Meehl (1970), Pearl's theory shows with unprecedented precision the fallacy of this approach: there are some variables that *must* be statistically controlled and others that *must not* be so controlled. In other words it is untrue that statistically controlling variables correlated with both putative cause and effect will either take us closer to the truth or do no harm; sometimes such "control" can take us *further* from the truth. I illustrate this insight and many others delivered by Pearl's theory in a number of realistic examples and a reanalysis of the Deary et al. (2008) data on intelligence and social liberalism. I also provide an extended discussion of why the techniques that have failed as tools of causal inference in so many other contexts have proved so robust in GWAS.

Pearl explains his motivation for writing his book as follows:

Ten years ago, when I began writing *Probabilistic Reasoning in Intelligent Systems* (1988), I was working within the empiricist tradition. In this tradition, probabilistic relationships constitute the foundations of human knowledge, whereas causality simply provides useful ways of abbreviating and organizing intricate patterns of probabilistic relationships. Today, my view is quite different. I now take causal relationships to be the fundamental building blocks both of physical reality and of human understanding of that reality, and I regard probabilistic relationships as but the surface phenomena of the causal machinery that underlies and propels our understanding of the world. (Pearl, 2009, pp. xv-xvi)

Such a philosophical commitment to causality over probability implies that there can be no "science of correlations." My hope is that the third paper convinces the reader that it is desirable and feasible for personality psychology to shed this label and convince its skeptics that its subject matter is indeed the causes of individual differences.

# **1 A Genome-Wide Association Study of 100+ Physical and Behavioral Traits Finds Few Loci of Large Effect**

James J. Lee<sup>1</sup>, Gregoire Borst<sup>2</sup>, Jonathan P. Beauchamp<sup>3</sup>, Daniel J. Benjamin<sup>4</sup>, Edward L. Glaeser<sup>3</sup>, Steven Pinker<sup>1</sup>, David I. Laibson<sup>3</sup>, Christopher F. Chabris<sup>5</sup>

**1 Department of Psychology, Harvard University, Cambridge, MA, USA**

**2 Groupe d’Imagerie Neurofonctionnelle du Developpement, Université Paris Descartes, Sorbonne Paris Cité, Paris, France**

**3 Department of Economics, Harvard University, Cambridge, MA, USA**

**4 Department of Economics, Cornell University, Ithaca, NY, USA**

**5 Department of Psychology, Union College, Schenectady, NY, USA**

## **Abstract**

We present the results of a genome-wide association study of over 100 human phenotypes, including body size, pigmentation, and many traits of interest to behavioral scientists. In a scan of over 660,000 single-nucleotide polymorphisms assayed in more than 400 participants, we replicate “positive control” associations reported in previous studies (most notably eye and hair color) but fail to detect appreciable signal for cognitive ability, personality, and the other behavioral traits. In particular, there is a conspicuous failure to replicate findings from previous gene-trait association studies of these traits that employed similar sample

sizes. Studies of twins and other kinships have shown that behavioral traits are highly heritable; our findings support the view that these traits resemble height or BMI in that their heritabilities are attributable to many loci of very small effect. Drawing on Fisher's geometric model of adaptation, we offer the hypothesis that the different evolutionary trajectories of the traits examined in our study account for their disparate genetic architectures.

## **1.1 Introduction**

Genome-wide association studies (GWAS) offer the potential to uncover much of a given trait's genetic architecture: the genomic locations, average effects, and allele frequencies of the DNA variants affecting the trait. The identification of a trait's genetic architecture, even if only fragmentary, should prove a great boon to scientists studying the basic biological mechanisms connecting genetic and phenotypic variation. In addition, whereas an individual's genome provides a partial blueprint for the development of the phenotype forward in time, our species' array of genomic data provides a partial record of our evolutionary history backward in time. Thus, knowledge of the genetic architecture may shed light not only on proximate biological mechanisms, but also the ultimate evolutionary forces that have shaped the commonality and diversity of humankind. These powerful motivations compel extending GWAS to behavioral traits of fundamental importance in differential and evolutionary psychology.

Here we present the results of a GWAS of over 100 human phenotypes, including body size, pigmentation, and many traits of interest to behavioral scientists. To our knowledge this study is the first to examine associations between a genome-wide panel of single-nucleotide polymorphisms (SNPs) and such a broad spectrum of phenotypes; in particular, most re-

ported gene-trait association studies of behavioral traits have singled out only a few candidate genes. We do not observe any novel and unambiguous SNP-trait associations at an appropriately stringent significance threshold. Since many of our measured phenotypes are known to be heritable (Plomin et al., 2008), the absence of strong associations in our data indicates that our traits of interest are affected in the main by numerous loci of small effect. Given the findings to date from GWAS of diseases and anthropometric traits (Manolio et al., 2009), this conclusion is perhaps unsurprising. In our Discussion we offer a novel extension of Fisher’s (1999) geometric model of adaptation, providing evolutionary rationales for (1) why a typical genetic architecture consists of many “infinitesimal” loci and (2) why exceptions to this trend can be found. We then discuss the implications of our findings for future association studies of behavioral traits.

## 1.2 Results

As can be seen in Table 1, we found at least marginal signal for all SNPs previously found to be associated with eye color, hair color, freckling, and skin color (Stokowski et al., 2007; Sulem et al., 2007, 2008; Sturm et al., 2008; Han et al., 2008; Eriksson et al., 2010)—except those reported by Liu et al. (2010) using digital quantification of eye color—and that were either present in our cleaned set of genotyped SNPs or represented by a proxy SNP with an  $r^2 > .6$ . Note that the effects of the intronic SNP rs12913832 in *HERC2* on eye and hair color were statistically significant at the stringent threshold appropriate for GWAS.

A meta-analysis has identified over 180 genomic regions containing a variant affecting height (Lango Allen et al., 2010). Due to the weak effect of any single variant, we did not replicate any of these loci at a stringent significance threshold. However, of the 94 loci either



Table 1: Association results for pigmentation phenotypes.

trait	reported SNP	proxy SNP	$r^2$	minor allele	sample MAF	HapMap MAF	effect size	$p$ -value	gene
eye darkness	rs12913832			A	.222	.208	.998	$2 \times 10^{-68}$	<i>HERC2</i>
eye darkness	rs12896399	rs1075830	.615	A	.460	.308	.167	.003	<i>SLC24A</i>
eye darkness	rs1393350			A	.266	.192	-.154	.02	<i>TYR</i>
eye darkness	rs1408799			T	.313	.300	.095	.11	<i>TYRP1</i>
hair darkness	rs12913832			A	.223	.208	.840	$1 \times 10^{-13}$	<i>HERC2</i>
hair darkness	rs12896399	rs1075830	.640	A	.460	.308	.372	$9 \times 10^{-5}$	<i>SLC24A4</i>
hair darkness	rs12821256			C	.095	.142	-.352	.03	<i>KITLG</i>
red hair	rs1805007			T	.076	.147	7.44	$2 \times 10^{-6}$	<i>MC1R</i>
red hair	rs1015362			T	.278	.233	.507	.09	<i>ASIP</i>
freckling	rs1805007			T	.076	.147	.613	$6 \times 10^{-6}$	<i>MC1R</i>
freckling	rs1042602			A	.346	.417	-.223	.005	<i>TYR</i>
freckling	rs2153271	rs1416742	.949	G	.384	.373	-.139	.07	<i>BNC2</i>
freckling	rs619865			A	.098	.108	.178	.15	<i>ASIP</i>
skin darkness	rs1805007			T	.076	.147	-.267	.005	<i>MC1R</i>
skin darkness	rs1042602			A	.346	.417	-.118	.03	<i>TYR</i>
skin darkness	rs619865			A	.098	.108	-.156	.07	<i>ASIP</i>

Eye darkness was reported on a 3-point scale. Hair darkness was recorded on 9-point scale. Red hair was recorded as a dichotomous trait, and its effect size is reported as an odds ratio. Freckling and skin darkness were recorded on 5-point scales. All effect sizes for non-dichotomous traits are reported as the expected change in trait value per each additional copy of the minor allele. All alleles are coded according to NCBI build 36 coordinates on the forward strand.

Table 2: Association results for physical phenotypes.

trait	reported SNP	proxy SNP	$r^2$	minor allele	sample MAF	HapMap MAF	effect size	$p$ -value	gene
standing height	rs7460090			C	.134	.117	-.188	.07	<i>SDR16C5</i>
standing height	rs237743			A	.231	.308	.175	.04	<i>ZNFX1</i>
standing height	rs6439167			T	.201	.183	-.191	.03	<i>C3orf47</i>
standing height	rs889014			T	.347	.375	-.124	.10	<i>BOD1</i>
standing height	rs7274811	rs3213183	.692	A	.304	.267	-.140	.07	<i>ZNF341</i>
standing height	rs7759938	rs369065	1	C	.332	.364	.172	.02	<i>LIN28B</i>
standing height	rs3764419	rs9890032	.982	G	.401	.375	-.188	.009	<i>ATAD5/RNF135</i>
standing height	rs3791675			T	.228	.275	-.305	$4 \times 10^{-4}$	<i>EFEMP1</i>
standing height	rs724016			G	.428	.483	.121	.10	<i>ZBTB38</i>
standing height	rs1351394	rs7968682	.983	T	.499	.517	-.120	.10	<i>HMGA2</i>
strength	rs1815739	rs540874	1	A	.428	.458	.252	.006	<i>ACTN3</i>

Effect sizes for height are reported in standard deviation units. Note that these effect sizes tend to be inflated because of the “winner’s curse.” Strength was reported on a 5-point scale.

present in our set of SNPs or represented by a proxy, 65 loci had estimated effects with the correct sign (binomial test  $p < 1 \times 10^{-4}$ ). There is also an enrichment of low  $p$ -values; whereas only nine or ten  $p$ -values less than .10 were expected under the null distribution, we observed 16 (binomial test  $p < .05$ ). These trends are consistent with most of these loci being true positives despite our inability to extract strong signal from them. A selection of the height variants showing marginal significance in our data is shown in Table 2, along with the nonsynonymous SNP rs1815739 in *ACTN3* known to affect athletic performance (MacArthur et al., 2007).

Another recent meta-analysis has identified 32 genomic regions containing a variant affecting body mass index (BMI) (Speliotes et al., 2010). The genetic architecture of BMI, relative to that of height, seems to be distributed among even more loci of small effect. In line with this trend, 11 of the 17 known BMI loci represented in our data had estimated effect sizes of the correct sign, but the wrong-signed loci were the most statistically significant.

Table 3 shows our results for a selection of SNPs previously reported to be associated with general cognitive ability (Payton et al., 2003; Plomin et al., 2004; Gosso et al., 2006; Zinkstock et al., 2007), personality (van den Oord et al., 2008; de Moor et al., 2011), working memory (Egan et al., 2001), and episodic memory (Papassotiropoulos et al., 2006). We observed little evidence of signal among these SNPs. In concordance with Need et al. (2008), replication of the association between *KIBRA* and episodic memory failed despite putative functional validation in the original study by both analysis of gene expression and fMRI. This suggests that most of the SNPs reported in earlier association studies of behavioral traits may represent either false positives or overestimates of the effect sizes. Applying a threshold of  $5 \times 10^{-8}$ , we did not observe any loci significantly associated with the traits in

Table 3: Association results for the behavioral phenotypes.

trait	reported SNP	proxy SNP	$r^2$	minor allele	sample MAF	HapMap MAF	effect size	$p$ -value	gene
general cognitive ability	rs2760118	rs7775073	.982	G	.316	.317	.062*	.42	<i>ALDH5A1</i>
general cognitive ability	rs324650			T	.464	.467	.026	.72	<i>CHRM2</i>
general cognitive ability	rs363050			G	.444	.475	-.027	.72	<i>SNAP-25</i>
general cognitive ability	rs17571	rs17834326	.781	A	.083	.083	-.051*	.70	<i>CTSD</i>
general cognitive ability	rs760761	rs2619545	1	C	.196	.192	-.033*	.72	<i>DTNBP1</i>
Conscientiousness	rs2576037	rs7233515	.879	A	.400	.408	-.038	.60	<i>KATNAL2</i>
Neuroticism	rs12883384			A	.410	.317	-.014*	.85	<i>MAMDC1</i>
paired-associate recognition	rs17070145			T	.338	.267	.065	.37	<i>KIBRA</i>
3-back accuracy	rs4680			A	.449	.517	.027	.72	<i>COMT</i>

Effect sizes are reported in sample standard deviation units. An asterisk indicates that the estimated effect in our study had a sign opposite to what had been previously reported.

Table 3.

We did find a significant association between political conservatism and rs10952668 (Table 4). This SNP lies in *LOC642355*, a pseudogene on chromosome 7. The SNP also showed an association with the highly correlated trait of Democrat vs Republican ( $\beta = .260$ ,  $p < .02$ ). We also observed a significant association between rs1402494 and gambling frequency; rs1402494 lies in a gene desert on chromosome 4. It happens that rs10952668 showed marginal evidence for association with the personality traits Openness ( $\beta = .142$ ,  $p < .06$ ) and Agreeableness ( $\beta = .130$ ,  $p < .08$ ). This raises the possibility that the association between political conservatism and rs10952668 is attributable to selection bias. Since to our knowledge this potential artifact has not been discussed in the genetic epidemiology literature (although it has parallels in the effects of natural selection on linkage disequilibrium), we discuss it at some length in the Materials and Methods. Upon the addition of general cognitive ability, Openness, Neuroticism, and Agreeableness as covariates in an attempt to control for selection bias, the association of rs10952668 and conservatism diminished and fell short of significance. The association of rs1402494 and gambling frequency appears robust against our attempts to control for selection bias, although it too fell short of significance after further adjustment. We conclude that both of these associations require replication in future studies.

The two SNP-trait associations in Table 4 were the only novel ones reaching the significance threshold  $5 \times 10^{-8}$  in our study. In summary, of all the traits in Table 5, only eye color, hair color, political conservatism, and gambling frequency yielded significant associations.

Table 4: **Novel association results for behavioral phenotypes.**

trait	reported SNP	minor allele	sample MAF	HapMap MAF	effect size	<i>p</i> -value
liberal vs conservative	rs10952668	T	.458	.392	.552 (.478)	$2 \times 10^{-8}$ ( $1 \times 10^{-6}$ )
gambling frequency	rs1402494	G	.206	.241	.278 (.276)	$3 \times 10^{-8}$ ( $6 \times 10^{-8}$ )

Liberal vs conservative was reported on a 7-point scale. Gambling frequency was reported on a 5-point scale. Effect sizes and *p*-values after adjustment for general cognitive ability, Openness, Neuroticism, and Agreeableness are given parenthetically. The effect estimates may be inflated as a result of the winner's curse.

### 1.3 Discussion

Given a significance threshold of  $5 \times 10^{-8}$ , our study had a power approaching .80 to detect any locus accounting for more than 10 percent of the variance in any particular trait. We retained reasonably good power (.12) for loci accounting for as little as 5 percent of the variance. The fact that we measured so many phenotypes implies that we would have obtained several hits if a large proportion of the phenotypes were indeed affected by such loci. Because we only obtained at most two new hits, however, loci with effects of this magnitude on the non-pigmentation traits in Table 5 must be quite uncommon. In agreement with Davis et al. (2010) and de Moor et al. (2011), we conclude that cognitive ability, personality dimensions, social attitudes, and most other traits of interest to behavioral scientists are affected by numerous loci of small effect. In this respect the behavioral traits in Table 5 resemble height and BMI rather than pigmentation.

There are at least two possible objections to the generalization that the genetic architecture of a typical quantitative trait consists of many loci of small effect. The first derives from studies of inbred strains of mice and *C. elegans* finding a number of closely linked loci with large phenotypic effects (Yazbek et al., 2010; Eric Evans, personal communication). However, the generalizability of these studies to outbreeding populations is highly uncertain. Given the small effective population sizes at which inbred strains are maintained over several generations, these strains might harbor genetic variants whose contributions to the variability of natural populations would be kept negligible by selection. It should be kept in mind that we probably all harbor very rare or unique variants with potentially large effects on some phenotype given the right genomic background, and geneticists studying an inbred line derived from forced brother-sister matings among the descendants of a few founders may well

detect a number of such variants. Thus, although analyses of inbred laboratory animals may be highly informative with respect to biological pathways, it is unclear whether they have much bearing on the genetic architecture of a typical quantitative trait in the human species.

The second possible objection is that many of the association signals picked up so far by GWAS are actually *synthetic associations* tagging one or more rare variants of large effect (Dickson et al., 2010; Wang et al., 2010; Goldstein, 2011b). If this objection is valid, then a relatively small number of genomic regions harboring multiple rare but very powerful variants may constitute a typical genetic architecture. This argument, however, has been thoroughly critiqued by Anderson et al. (2011) and Wray et al. (2011). The main points are summarized as follows:

1. If an association signal is explained by (possibly multiple) rare variants of large effect, then by far most association signals should arise from the rarer variants on SNP chips. This is because rare variants are tagged better by another rare variant than by a common variant. But what we observe in GWAS data is that although there are more signals from rare variants, there are still too many signals from common variants for a “rare-only” model.
2. For a very common variant to be significantly associated with a trait when the causal variant is actually rare, that rare variant must have an enormous effect that would have been detected in linkage studies. GWAS results do not align well with putative linkage signals, however, and so it seems that the GWAS signals at common tag variants are in fact mostly attributable to common causal variants.
3. If most association signals are due to rare variants of large effect—to be concrete, if



the phenotype is height, about 2.5 inches per minor allele with frequency .005—then it is possible for about a hundred loci, located in a rather small fraction of the genome, to account for all of the trait variance. But in the case of height there have been 180 loci identified so far and strong reason to believe that the number will continue to climb (Lango Allen et al., 2010; Turchin, 2011). Also, it has been shown that height loci are scattered throughout the genome at roughly constant density (the density being somewhat higher in genic regions) (Visscher et al., 2007; Yang et al., 2011).

4. The sign of a marker-trait association signal is often replicated in different racial groups. This strongly implies that the causal variant is old, its origin by mutation preceding the worldwide dispersal of *Homo sapiens*. Older variants must tend to be common.

For these reasons and others, I am convinced that most GWAS association signals are not synthetic and thus see no reason to modify my expectation that a typical behavioral trait will be affected by thousands of variants both common and rare.

The contrast between pigmentation and the other phenotypes examined in this study is quite striking (Tables 1–3). A question that naturally arises is whether any theoretical principles might explain why a diffuse polygenic architecture should be typical for a quantitative trait and also what accounts for the known exceptions to this trend. Perhaps the simplest possible explanation invokes the length of the causal chain from genetic to phenotypic variation. For example, variation in pigmentation arises from the number of melanosomes produced, the type of melanin synthesized, and the size and shape of the melanosomes (Sturm, 2009). It is plausible that these biochemical differences follow directly from changes in the compo-

sition or regulation of gene products. In contrast, changes at the molecular and cellular level may be somewhat remote from any ultimate changes in even a physical phenotype such as BMI. Consider that BMI may depend on what a person likes to eat, how often he eats, how much he exercises, and a host of other complex behaviors. Similarly, given the abstractness of psychological attributes such as cognitive ability, conscientiousness, religiosity, and the like, we might expect any single genetic variant affecting such an attribute to contribute little variability relative to the total causal background.

Another possible explanation is the differential action of natural selection. The essential features of Fisher's (1999) well-known geometric model of adaptation are captured in the two-dimensional phenomic space of Figure 1.  $A$  represents the current mean phenotype of the species, while  $O$  represents the optimum favored by natural selection. We imagine that  $A$  and  $O$  no longer coincide because of an abrupt environmental change demanding a different value of trait 1. The effect of fixing a new mutation in this model corresponds to adding a vector of random direction to the population's current position at  $A$ . This feature of the model captures two key observations: (1) mutations have no inherent tendency to increase the fitness of their bearers, and (2) any single mutation may affect several distinct traits.

The interior of the circle contains all new phenotypes that would result in an increased level of adaptation. It is immediately clear that selection forbids the fixation of any mutation whose magnitude exceeds the diameter of the circle. In general, mutations become more likely to be beneficial as their magnitudes decrease. For this reason Fisher argued that mutations of large effect are relatively unimportant in evolution.

The conformity of these statistical requirements with common experience will be perceived by comparison with the mechanical adaptation of an instrument,

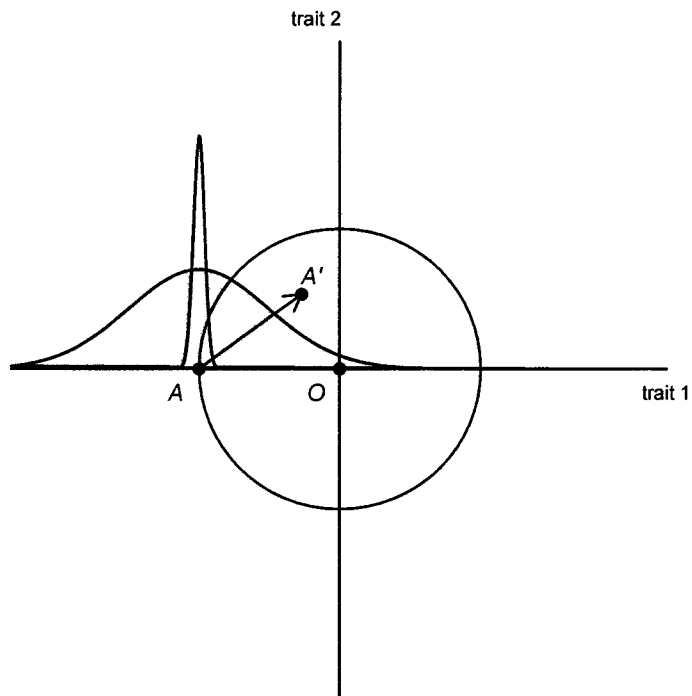


Figure 1: **Fisher's geometric model of adaptation.**  $A$  is the current mean phenotype of the population,  $A'$  is the mean phenotype that would result if the mutation denoted by the arrow were to be instantly fixed, and  $O$  is the new optimum favored by natural selection.

such as a microscope, when adjusted for distinct vision. If we imagine a derangement of the system by moving a little of the lenses, either longitudinally or transversely, or by twisting through an angle, by altering the refractive index and transparency of the different components, or the curvature, or the polish of the interfaces, it is sufficiently obvious that any large derangement will have a very small probability of improving the adjustment, while in the case of alterations much less than the smallest of those intentionally effected by the maker or the operator, the chance of improvement should be almost exactly half. (Fisher, 1999, pp. 40-41)

We now expand Fisher's argument to address the puzzle raised by the contrast between pigmentation and the other traits in Table 5.

If the distance between  $A$  and  $O$  in Figure 1 is very large, then it is possible that a mutation with a sizable projection on trait 1 will be favored by selection. Suppose that trait 1 was previously under strong stabilizing selection and thus has negligible genetic variation at the time of the environmental shift (corresponding to a tight clustering of phenotypes around  $A$ ). Since the rate of the approach to the optimum by standing genetic variation is bounded above by trait 1's heritability (Lande, 1979), a population with no variability in trait 1 would be fortunate to fix a mutation taking it to  $A'$ .

But now suppose that stabilizing selection on trait 1 was much weaker, permitting the buildup of substantial genetic variation (a wide scatter of points around  $A$ ). In this case it becomes much less probable that a mutation of large effect will become common as a result of positive selection. Standing genetic variation (as well as environmental variation) swamps the fitness effect of a new mutation in essentially random noise and thus retards its progress away from frequency zero. At the same time, standing genetic variation enables the population to advance toward  $O$  while the mutation is struggling to escape from the boundary. If  $O$  lies within the current range of genetic variation (as is the case for the more dispersed popu-

lation in Figure 1) and selection is even moderately strong, then the population mean shifts from  $A$  to  $O$  in just a few generations. *En route* the diameter of the circle bounding all points of higher adaptation continuously shrinks. Once the magnitude of the mutation that would have taken the population to  $A'$  exceeds the diameter of the circle, the mutation is disfavored and very likely absorbed back at frequency zero.

What kinds of genetic variants contribute to standing variation? Even under weak stabilizing selection, variants of large effect are more easily “seen” by selection and consequently kept at a low minor allele frequency (MAF) (Wright, 1938; Hastings, 1990; Eyre-Walker, 2010). This implies that any common variants contributing to standing genetic variation will typically be small in effect. Thus, we might expect many loci of small effect to be the typical genetic architecture underlying a quantitative trait.

In retrospect, these evolutionary considerations may account for the pattern evident in Tables 1–3. After the loss of body hair in our lineage, pigmentation came under strong stabilizing selection in our ancestors, who were in great need of protection from the African sun. More recently, the out-of-Africa migrants ancestral to Europeans and East Asians experienced a sudden and drastic shift in the optimal level of pigmentation—perhaps because of the need to sustain cutaneous synthesis of vitamin D in northern climates (Jablonski & Chaplin, 2010), although others have implicated sexual selection or as-yet unidentified evolutionary pressures (Cavalli-Sforza et al., 1994; Frost, 2006; Cochran & Harpending, 2009). In any event the result was that several depigmenting mutations of large effect increased very rapidly in frequency (Rogers et al., 2004; Williamson et al., 2007; Pickrell et al., 2009). Table 1 lists those mutations that have not yet reached fixation and are thus still polymorphic in Europeans. On the other hand, phenotypes such as height, BMI, and the behavioral traits

in Table 5 have probably always been quite variable in human populations. Indeed, evolutionary game theory has established theoretical rationales for the persistence of multiple behavioral phenotypes (e.g., hawks and doves) in the same population (Maynard Smith & Price, 1973; Penke et al., 2007). Even if selection has acted on these traits since the dispersal of our species from Africa, the new optimums could have been quickly reached by small shifts in allele frequency at many minor loci, leaving any major mutants at the low MAF determined by the interaction of mutation, drift, and stabilizing selection (Kimura, 1983). The result of these dynamics would be the observed absence of common variants with large effects.

Note that our explanation appealing to the length of the causal chain between genotype and phenotype can easily be rephrased to say that different traits have different distributions of mutational effects. This would be the kind of “nonadaptive” explanation for an observed genetic phenomenon advocated by Lynch (2007), whereas our hypothesis regarding the influence of natural selection would be an “adaptive” explanation. We feel that our adaptive and nonadaptive hypotheses are not in conflict, and probably both contribute to the differences in genetic architectures across quantitative traits.

Our two proposals for explaining the pattern in Tables 1–3 lead to the following suggestions for future GWAS of behavioral traits. First, in order to understand the various steps in the causal chain between genetic and phenotypic variation, attempts should be made to narrow the chasm from both sides. This requires the validation of endophenotypes lying closer on the causal chain to genetic variation than the phenotypes of primary interest. Second, researchers seeking variants of large effect should study populations where directional selection may have recently produced a phenotypic change that is large relative to the initial

standing variation. Recent studies of altitude adaptation in Tibetans exemplify both of these suggestions (Simonson et al., 2010; Yi et al., 2010; Beall et al., 2010). The genes associated with red blood cell count and hemoglobin concentration in these studies would have been more difficult to identify if the phenotype had been characterized at a level as abstract as “altitude tolerance.” Moreover, the recent and rapid divergence of Han Chinese and Tibetans in altitude tolerance after the latter began to occupy a highland environment was plausibly driven by a selection differential large enough to pull variants of large effect away from the boundary of frequency zero.

As for traits experiencing more typical evolutionary histories, one promising approach to collecting the required large samples has been the burgeoning field of personal genomics, in which a large base of volunteers or consumers provide genotype and phenotype information (Dolgin, 2010; Lunshof et al., 2010; Eriksson et al., 2010). In such studies it will be important to check for selection bias to the extent possible. Although a trait such as asparagus anosmia plausibly has little influence on appearing in a research study, other phenotypes of interest in personal genomics are likely to be causes of participation in personal genomics itself. For example, an individual with a liability to a particular disease may be strongly motivated to participate in a personal genomics study for reasons of self-interest or altruism. We suspect that our findings of elevated cognitive ability and intellectual openness among research volunteers will generalize to future studies, and therefore it may be prudent to collect highly reliable measurements of these traits in all participant-driven GWAS. If the sample distributions of these traits appear unusual, then it may be useful to annotate reported associations to indicate that they may have arisen in a conditional background. Traditional epidemiological studies that attempt to minimize the impact of personal characteristics on

study participation will remain an important complement to volunteer- and consumer-driven approaches.

In summary, we find very few loci of large effect associated with non-pigmentation traits, including many traits of great theoretical interest to behavioral scientists. Two substantial points emerge from our analysis:

1. The evolutionary history of a trait is intimately intertwined with its present genetic architecture. This implies a bidirectional impact: just as the discovery of gene-trait associations can illuminate the evolution of our species, the realized evolutionary process affects what associations we can most readily discover. In particular, stabilizing selection of middling strength, which permits a substantial background of weak or rare variants, supplies the fuel for polygenic adaptation and may obviate the need for mutants of large effect upon a sudden environmental change.
2. Psychological traits of interest to researchers are themselves very likely to be causes of participation in scientific research, which raises the potential of spurious associations arising from self-selection.

These points will remain of relevance as gene-trait association studies of behavioral traits proceed under various approaches.

## **1.4 Materials and Methods**

### **1.4.1 Ethics Statement**

This study was conducted according to the principles expressed in the Declaration of Helsinki. The participants in this study all provided written informed consent. All proce-



dures and the consent form were approved by the Harvard Committee on the Use of Human Subjects in Research.

#### **1.4.2 Participants**

Participants were recruited and phenotyped at two sites: Cambridge, MA, and Schenectady, NY. Participants were recruited through paper flyers posted at various sites, advertisements placed on Craigslist, and the Department of Psychology Study Pool at Harvard University.

Participants were directed to a SurveyMonkey questionnaire that included items regarding age, medical history, and grandparental ethnicity. Any participants who reported an age outside the range 18 to 45 or a history of bipolar disorder, schizophrenia, or severe head trauma were excluded from followup. To control for ancestral confounding of genotypes and trait levels (Campbell et al., 2005), we selected a sample of predominantly Western European ancestry.

We phenotyped 451 participants. During the phenotyping some participants reported discrepant or more detailed self-reports regarding the eligibility criteria that disqualified them. These participants were phenotyped but not genotyped, leaving 419 participants with complete genetic and phenotypic data.

#### **Measures**

Table 5: Phenotypes measured in the study

phenotype	mode	scale
3-back accuracy	computer	quantitative (N)
3-back RT	computer	quantitative (N)
acne severity as adolescent	self-report	polytomous
acne severity as adult	self-report	polytomous
acne severity overall	self-report	polytomous
alcohol consumption frequency (last 12 months)	self-report	polytomous
alcohol drinks per drinking occasion	self-report	quantitative
alcohol total drinks in last year	self-report	quantitative (N)
allergic to animals	self-report	dichotomous
allergic to drugs	self-report	dichotomous
allergic to food	self-report	dichotomous
allergies (any)	self-report	dichotomous
anticipated remaining life expectancy	self-report	quantitative (N)
asthma as adult	self-report	dichotomous
asthma as child	self-report	dichotomous
athleticism	self-report	polytomous
attitude toward abortion on demand	self-report	polytomous
attitude toward alcohol	self-report	polytomous
attitude toward attention-drawing clothes	self-report	polytomous

*Figure 5 continued on next page*

phenotype	mode	scale
attitude toward being the center of attention	self-report	polytomous
attitude toward being the leader of groups	self-report	polytomous
attitude toward big parties	self-report	polytomous
attitude toward capitalism	self-report	polytomous
attitude toward castration as sex crime punishment	self-report	polytomous
attitude toward death penalty for murder	self-report	polytomous
attitude toward doing athletic activities	self-report	polytomous
attitude toward dressing well at all times	self-report	polytomous
attitude toward education	self-report	polytomous
attitude toward exercising	self-report	polytomous
attitude toward getting along well with others	self-report	polytomous
attitude toward illegal drugs	self-report	polytomous
attitude toward legalized gambling	self-report	polytomous
attitude toward loud music	self-report	polytomous
attitude toward making racial discrimination illegal	self-report	polytomous
attitude toward open-door immigration	self-report	polytomous
attitude toward organized religion	self-report	polytomous
attitude toward playing chess	self-report	polytomous
attitude toward playing organized sports	self-report	polytomous
attitude toward public speaking	self-report	polytomous
attitude toward reading books	self-report	polytomous

*Figure 5 continued on next page*

phenotype	mode	scale
attitude toward rollercoaster rides	self-report	polytomous
attitude toward smoking	self-report	polytomous
attitude toward voluntary euthanasia	self-report	polytomous
back pain	self-report	dichotomous
BIS inattention	self-report	quantitative (N)
BIS general	self-report	quantitative (N)
BIS motor	self-report	quantitative (N)
BIS nonplanning	self-report	quantitative (N)
body mass index	measured	quantitative (N)
body type (scrawny to obese)	self-report	polytomous
caffeine mg per day	self-report	quantitative
CFMT	computer	quantitative (N)
cigarette packs per day	self-report	polytomous
cleft chin	self-report	dichotomous
coffee cups per day	self-report	polytomous
corrective lenses needed currently	self-report	dichotomous
corrective lenses needed at any time	self-report	dichotomous
curl tongue	self-report	dichotomous
Democrat vs. Republican	self-report	polytomous
dental braces worn (ever)	self-report	dichotomous
dental braces worn or needed (ever)	self-report	dichotomous

*Figure 5 continued on next page*

phenotype	mode	scale
dictator game	self-report	dichotomous
dimples	self-report	dichotomous
discounting	self-report	quantitative (N)
drink alcohol (ever)	self-report	dichotomous
earlobes free (vs. hanging)	self-report	dichotomous
evening person	self-report	dichotomous
exercise amount per week	self-report	polytomous
exercise intensity	self-report	polytomous
exercise regularly	self-report	dichotomous
eye color	self-report	polytomous
facial hair color	self-report	polytomous
facial hair color (red vs. not red)	self-report	dichotomous
farsighted	self-report	dichotomous
first toe longer than second toe	self-report	dichotomous
floss teeth regularly	self-report	dichotomous
freckles on face	self-report	polytomous
gambling frequency	self-report	polytomous
general cognitive ability		quantitative (N)
hair color	self-report	polytomous
hair color (red vs. not red)	self-report	dichotomous
hair curliness	self-report	polytomous

*Figure 5 continued on next page*

phenotype	mode	scale
hair on middle segment of any finger	self-report	dichotomous
happiness sumscore	self-report	quantitative (N)
hay fever	self-report	dichotomous
heterosexual	self-report	dichotomous
hitchhiker's thumb	self-report	dichotomous
hours of sleep average	self-report	quantitative
hours of sleep last night	self-report	quantitative
illegal drug use	self-report	polytomous
inattentional blindness	computer	dichotomous
in-person contact with family or very close friends	self-report	dichotomous
last doctor's appointment for checkup	self-report	polytomous
liberal vs conservative	self-report	polytomous
loss aversion	self-report	quantitative
MAB Arithmetic	paper	quantitative (N)
MAB Similarities	paper	quantitative (N)
MAB Vocabulary	paper	quantitative (N)
memory problems	self-report	dichotomous
migraines at any time	self-report	dichotomous
migraine frequency	self-report	polytomous
migraine within last 12 months	self-report	dichotomous
morning person	self-report	dichotomous

*Figure 5 continued on next page*

phenotype	mode	scale
multivitamin supplement	self-report	dichotomous
nearsighted	self-report	dichotomous
NEO Agreeableness	self-report	quantitative (N)
NEO Conscientiousness	self-report	quantitative (N)
NEO Extraversion	self-report	quantitative (N)
NEO Neuroticism	self-report	quantitative (N)
NEO Openness	self-report	quantitative (N)
paired-associate recognition	computer	quantitative (N)
percentproxe of income saved over last 3 years	self-report	quantitative
physical attractiveness	self-report	polytomous
quality of sleep	self-report	polytomous
RAPM	computer	quantitative (N)
religiosity	self-report	quantitative
right-handed	self-report	dichotomous
risk aversion	self-report	quantitative (N)
seat belt use	self-report	polytomous
shape span accuracy	computer	quantitative (N)
shape span response time	computer	quantitative (N)
sitting height	measured	quantitative (N)
skin color and sun exposure response	self-report	polytomous
SMMR accuracy	computer	quantitative (N)

*Figure 5 continued on next page*

phenotype	mode	scale
SMMR response time	computer	quantitative (N)
smoked cigarette (ever)	self-report	dichotomous
soda cups per day	self-report	polytomous
spatial span accuracy	computer	quantitative (N)
spatial span response time	computer	quantitative (N)
SRTT accuracy	computer	quantitative
SRTT overall RT	computer	quantitative (N)
SRTT improvement in RT	computer	quantitative (N)
standing height	measured	quantitative (N)
strength	self-report	polytomous
stress level within last 12 months	self-report	polytomous
sunscreen or protective clothing use	self-report	polytomous
tea cups per day	self-report	polytomous
time woke up this morning	self-report	quantitative (N)
tobacco use frequency (current)	self-report	polytomous
tobacco user (current)	self-report	dichotomous
tobacco user (ever)	self-report	dichotomous
unprotected sex	self-report	polytomous
utilitarianism	self-report	quantitative (N)
verbal fluency	audio	quantitative (N)
vision quality (uncorrected)	self-report	polytomous

*Figure 5 continued on next page*



phenotype	mode	scale
VVIQ	self-report	quantitative (N)
weight	measured	quantitative (N)
weight (maximum)	self-report	quantitative (N)
widow's peak	self-report	dichotomous

Table 5 lists all measured phenotypes.

Any phenotype measured in *paper* mode was administered as a traditional paper-and-pencil test. *Self-report* refers to questionnaire data recorded either on paper forms or a SurveyMonkey questionnaire. Phenotypes measured in *computer* mode were implemented as PsyScope tasks requiring participants to provide keyboard input. A *measured* phenotype was directly measured by an experimenter using either a measuring tape or a bathroom scale. *Audio* refers to sound-recorded data that was later transcribed and coded.

Any variable assuming more than ten values was regarded as quantitative rather than polytomous (ordered categorical). A parenthetical N in Table 5 indicates that we were able to remove sex differences in mean and variance from a quantitative variable and then use a quantile transformation to render the resulting scores normally distributed. These transformations should increase statistical power to detect genetic associations for traits showing sex differences and also the accuracy of *p*-values.

We now describe some behavioral phenotypes whose labels in Table 5 are relatively uninformative.

- *3-back*. Participants viewed a succession of words, each new word appearing every 2.36 s. Participants were instructed to indicate as quickly and accurately as possible

whether each word matched the word seen three items previously. This task has often been employed as an indicator of working memory capacity (Gray et al., 2003).

- *Barratt Impulsiveness Scale (BIS)*. This self-report has been found to measure three distinct factors (inattention, motor impulsiveness, lack of planning) (Patton et al., 1995). We used the sum of these three factor scores as a measure of this self-report's general factor.
- *Cambridge Face Memory Test (CFMT)*. Participants were shown six target human faces and then tested with a forced-choice item consisting of three faces, one of which was a target. This test has been shown to be a sensitive measure of prosopagnosia (specific deficit in recognizing other people by their facial features) and also normal variability in the ability to recognize faces (Duchaine & Nakayama, 2006; Wilmer et al., 2010).
- *Dictator game*. Each participant was asked to imagine being randomly and anonymously paired with another participant. The participant was then asked to allocate ten dollars between the members of the pair. How much of the ten dollars each participant is willing to give away to the other person in this task has been taken as a measure of the participant's heritable altruistic tendencies (Knafo et al., 2008; Cesarini et al., 2009). Because the distribution of allocation was almost bimodal, nearly all participants giving away either zero or five dollars, we treated this phenotype as dichotomous; all participants who gave anything at all were given the higher score.
- *Discounting*. Participants were presented a set of choices between smaller prompt

rewards and larger delayed rewards. Discount rates inferred in this way have been found to be associated with substance abuse and other outcomes (Chabris et al., 2008).

- *General cognitive ability.* We combined the following indicators into a standardized cognitive ability composite: (1) the short form of Raven's Advanced Progressive Matrices proposed by Bors and Stokes (1998); (2) the Arithmetic, Similarities, and Vocabulary subtests of the Multidimensional Aptitude Battery (MAB); and (3) accuracy on a forced-choice version of the Shepard-Metzler Mental Rotations (SMMR).
- *Inattentional blindness.* Participants watched a video of two teams of three players, one team wearing white shirts and the other wearing black shirts, who moved around erratically in an open area. The passes were either bounce passes or aerial passes; players would also dribble the ball, wave their arms, and make other movements. After about 45s, a research assistant wearing a gorilla costume that fully covered her body walked through the action. A surprising proportion of participants report not seeing the gorilla at all (Simons & Chabris, 1999). The causes of the individual differences in this task are unknown. We treated any participant who reported having seen or heard of this task previously as a missing data point.
- *Loss aversion.* Participants were presented with a set of choices between (1) receiving nothing or (2) a 50% chance of gaining an amount  $x$  and a 50% chance of losing an amount  $y$ .
- *NEO Five-Factor Inventory.* A 60-item self-report with 12 items measuring each of the following personality factors: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness.

- *Paired-associate recognition.* After studying a series of arbitrary paired words, participants were given a multiple-choice test of cue recognition.
- *Religiosity.* We administered the religiousness scale employed by Koenig et al. (2005).
- *Risk aversion.* Participants were presented with a set of choices between (1) a 100% chance of receiving an amount  $x$  or (2) a 50% chance of receiving an amount  $y > x$  and a 50% chance of receiving nothing.
- *Shape span.* In the study phase, participants were presented a series of irregular shapes one at a time. In the test phase, participants had to press one of two keys in response to further presentations of irregular shapes, depending on whether each shape was a member of the set presented in the study phase.
- *Social attitudes.* Items asking for attitudes toward abortion on demand, alcohol, and so forth were taken from (Olson et al., 2001). Because the factor model postulated by these authors did not fit our data well, we analyzed each item separately.
- *Spatial span.* In the study phase, participants viewed a circular array of gray dots. Several of the dots briefly turned black, one at a time. In the test phase, participants continued to view dots turning black and had to indicate by pressing one of two keys whether a dot had also turned black during the study phase.
- *Serial Reaction Time Task (SRTT).* Participants viewed a linear array of four squares. During each trial a black diamond briefly appeared in one of the squares, and in response participants had to press one of four corresponding keys, using the pinky, ring, middle, and index fingers of the preferred hand. Unbeknownst to the participants, a

fixed subsequence of the stimuli appeared repeatedly throughout the task, alternating with runs of stimuli chosen at random. Reaction time (RT) tends to decrease with each successive presentation of the repeating subsequence, although most participants do not consciously notice the repetition. The mean difference in RT between the repeating stimuli and the random stimuli was taken as a measure of implicit motor learning.

- *Utilitarianism.* Participants were presented with a set of moral dilemmas in which participants rated on a 1-to-5 scale the appropriateness of a “utilitarian” response (Greene et al., 2001). A typical item stem: “You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. *Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?*”
- *Verbal fluency.* Participants were given one minute to utter as many distinct words as possible beginning with a certain letter. Person names, places, and numbers were not counted. The letters F, A, and S were used. The counts of the uttered words beginning with these letters appeared to be tau-equivalent indicators of a common factor after standardization.
- *Vividness of Visual Imagery Questionnaire (VVIQ).* Participants were told to visualize certain scenes or persons and rate the vividness of distinct aspects of the mental image

(Marks, 1973).

### 1.4.3 Procedure

Participants found to be eligible for the study after prescreening were invited to one of our two labs for a phenotyping session lasting typically from three to four hours. Participants gave informed consent after the nature of the procedure had been fully explained to them. At the end of the session, each participant rolled a six-sided die. If the participant rolled a six, then a randomly chosen response to an item in one of the behavioral-economic tasks (discounting, dictator game, loss aversion, risk aversion) was fulfilled for that participant. For example, suppose that the chosen item was a discounting item. If the participant expressed a preference for  $x$  dollars 30 days from now over  $y$  dollars 60 days from now, then the participant was written a check for  $x$  dollars dated 30 days from the date of the phenotyping session. Any losses suffered in the loss aversion task came out of five dollars given to each participant at the beginning of the session. This five dollars was given in addition to the advertised 50-dollar compensation. Participants were informed at the outset of the phenotyping session that their choices in the behavioral-economic tasks might actually be fulfilled.

At two points during the phenotyping session, participants provided samples by washing their mouths with 10 ml of Scope mouthwash. Samples were stored either in a freezer at  $-20^{\circ}\text{C}$  or in packed dry ice until DNA extraction. Genomic DNA was extracted using a QIAamp DNA Blood Mini Kit according to the manufacturer's recommended protocol.

Genomic DNA samples normalized to  $50\text{ ng}/\mu\text{l}$  were genotyped at either Stanford Genome Technology Center (SGTC) or Expression Analysis (EA) in Durham, North Carolina using the Affymetrix Genome-Wide Human SNP Array 6.0 in four batches. SNP geno-

types were called using the Birdseed v2 algorithm applied to each batch individually. The median call rate before application of quality-control (QC) criteria was 99.64%. Between-batch reproducibility was assessed by genotyping both samples provided by each of two participants. Average genotype concordance between replicates was 99.7%.

We used the program PLINK for data cleaning and analyses (Purcell et al., 2007). Our QC criteria excluded all participants missing more than 7% of their genotypic data, all SNPs with minor allele frequency (MAF) less than .05, all SNPs deviating from Hardy-Weinberg equilibrium at a significance threshold of  $5 \times 10^{-8}$ , and all SNPs missing more than 5% of their calls. We then computed the principal components (PCs) of the resulting genotype matrix with the program EIGENSTRAT (Price et al., 2006). All participants more than six standard units from the origin on any of the top 10 PCs were iteratively excluded. After application of all QCs, the final cleaned dataset included 401 individuals and 661,107 SNPs.

Nine statistically significant PCs at a significance threshold of .05 were found. The PCs corresponding to the fourth and fifth largest eigenvalues weakly distinguished the two genotyping laboratories, despite the application of our earlier QCs. The first, second, third, and sixth PCs were significantly correlated with the geographical distance of grandparental origin from England. The seventh PC tended to spread out individuals reporting non-British grandparents, whereas the eighth PC tended to separate those reporting two or more British grandparents from those reporting one or none. The ninth PC tended to spread out individuals reporting British grandparents, perhaps reflecting structure within Britain. We included all nine significant PCs as covariates in the tests for SNP-trait association.

The MAB subtests were scored according to the instructions in the manual (Jackson, 1998). Factor analyses of the BIS, the NEO, religiosity, utilitarianism, verbal fluency, and

the VVIQ resulted in solutions with nonzero uniquenesses. For these phenotypes we estimated factor scores by Bartlett’s method, which is maximum likelihood (ML) if the uniquenesses are normally distributed. A few participants were missing some data as a result of omissions, photocopying errors, computer failures, and the like. Participants were given factor scores if they responded to more than half of a scale’s indicators. We used the OpenMx package in R to perform all factor analyses (R Development Core Team, 2010).

Parameters describing the responses of each participant during the behavioral-economic tasks were estimated by ML. For example, an “interest rate” for discounting utility flows over time was estimated for each person and used as the measurement for the discounting task.

Linear regression was performed to test for purely additive association between SNPs and all polytomous and continuous traits. Logistic regression was performed for dichotomous traits. We chose a significance threshold of  $5 \times 10^{-8}$  for declaring a SNP-trait association to reflect a causal effect of the marker or a linked variant (McCarthy et al., 2008). Under a frequentist approach aiming to minimize the chance that even a single declared “hit” is a false positive, the large number of examined traits requires an even more stringent threshold. However, we favor the quasi-Bayesian justification for the strict GWAS significance threshold given by the Wellcome Trust Case Control Consortium (2007). This approach recognizes that a given significance threshold maintains a constant ratio of true to false positives as the number of markers and traits increases—so long as statistical power and prior probabilities do not change. Thus, if a given threshold has already been shown to produce an acceptable ratio of true to false positives, then it is reasonable to expect that *most* declared hits will be true positives as a study tests more hypotheses—even if it becomes virtually certain that



*some* of the declared hits are false positives. The appeal of this reasoning is that it does not mandate prohibitive significance thresholds as association studies begin to exploit whole-genome sequencing and the measurement of multiple phenotypes.

For any SNP showing an association with a trait at the significance threshold  $5 \times 10^{-8}$ , we reran PLINK with our cognitive ability composite and NEO Openness, Neuroticism, and Agreeableness factor scores as additional covariates in an effort to control for selection bias (Pearl, 2009). To illustrate what we mean by selection bias, suppose that whether our driveway is wet is affected by whether it rained last night and whether our sprinkler was activated (Figure 2a). Suppose also that the two causal variables are independent; that is, when it rains is not associated with when the sprinkler turns on. If we only examine the pavement on mornings when it is wet, however, then the two causal variables become negatively correlated. For instance, if we see that the pavement is wet *and* know that it did not rain last night, then we can be fairly confident that the sprinkler was in fact activated. The basic principle emerging from this example is that conditioning on the common effect of multiple causes can introduce an association among the causes where none exists marginally.

This same principle applies in GWAS. Suppose that higher levels of both traits 1 and 2 are causes of participation in our study (Figure 2b). Then we will find any gene affecting trait 2 to be associated with trait 1, even if trait 1 is not at all affected by genetic variation. Controlling for the other traits affecting participation is not fully satisfactory, even if we know what these traits are. The trait of interest may itself be connected to the other traits in a complex causal graph, and therefore conditioning on the other traits may introduce further bias. In all likelihood, however, conditioning on traits that may affect study participation is a conservative procedure.

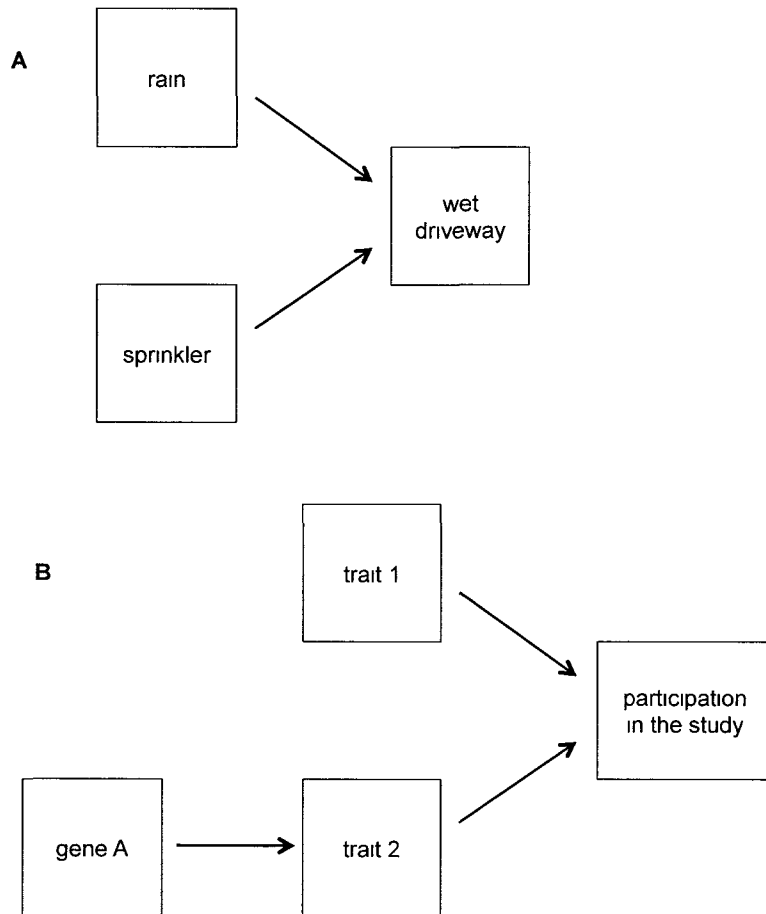


Figure 2: **Examples of directed acyclic graphs containing a collider (the common effect of two or more causes).** Conditioning on a collider alters the covariation among the causes; for example, two marginally independent causes can become negatively correlated.

**Table 6: Characteristics of the sample.**

trait	mean	SD
age	25.2	6.44
sex	67.6% female	
MAB Arithmetic	.797 (0)	.836 (1)
MAB Similarities	1.054 (0)	.601 (1)
MAB Vocabulary	1.386 (0)	.891 (1)
NEO Neuroticism (college, female)	21.90 (25.83)	8.38 (7.59)
NEO Neuroticism (adult, female)	18.71 (20.54)	9.13 (7.61)
NEO Neuroticism (college, male)	18.53 (22.49)	10.04 (7.92)
NEO Neuroticism (adult, male)	18.84 (17.60)	10.46 (8.61)
NEO Extraversion (college, female)	30.10 (31.27)	6.89 (5.64)
NEO Extraversion (adult, female)	29.19 (28.16)	7.55 (5.82)
NEO Extraversion (college, male)	29.08 (29.22)	6.10 (5.97)
NEO Extraversion (adult, male)	29.70 (27.22)	8.64 (5.85)
NEO Openness (college, female)	34.02 (27.94)	6.57 (5.72)
NEO Openness (adult, female)	34.42 (26.98)	5.57 (5.87)
NEO Openness (college, male)	31.79 (27.62)	6.57 (6.08)
NEO Openness (adult, male)	31.36 (27.09)	7.04 (5.82)
NEO Agreeableness (college, female)	33.80 (31.00)	5.51 (5.33)
NEO Agreeableness (adult, female)	34.42 (33.76)	4.71 (4.74)
NEO Agreeableness (college, male)	31.46 (28.76)	6.05 (5.24)
NEO Agreeableness (adult, male)	32.00 (31.93)	5.70 (5.03)
NEO Conscientiousness (college, female)	33.64 (31.02)	7.40 (6.53)
NEO Conscientiousness (adult, female)	32.29 (35.04)	7.15 (5.78)
NEO Conscientiousness (college, male)	30.17 (30.21)	6.54 (7.19)
NEO Conscientiousness (adult, male)	33.33 (34.10)	8.04 (5.95)

The summary statistics reported in the respective manuals are given in parentheses next to the corresponding sample statistics. The MAB scores were scaled as standard normal using the tables in Jackson (1998). The NEO summary statistics were calculated for participants between the ages of 18 and 22 for purposes of comparison with the college norms in Costa and McCrae (1992) and for participants age 30 and over for comparison with the adult norms.

Table 6 gives the sample statistics for the MAB and NEO, two instruments with detailed population norms. Our participants show much higher MAB means and smaller standard deviations than the norming samples, suggesting that cognitively able individuals were more likely to participate in the study. The relationship between the NEO personality traits and study participation appears to be rather complex. Our study participants show conspicuously higher levels of Openness than the norming samples. The trait of Openness is defined by a willingness to examine new ideas and try new activities, and thus it is quite plausible that higher levels of this trait may be a cause of participation in scientific research. Our study participants also show consistently lower levels of Neuroticism and higher levels of Agreeableness. Interestingly, our study participants show a pronounced tendency to be more variable than the norming samples, although this trend may be due partly to the fact that individuals with higher measured values of cognitive ability are more variable in their responses to personality questionnaires (Aitken Harris et al., 2005).

A reasonable attempt to control for selection bias thus appears to be using general cognitive ability, Openness, Neuroticism, and Agreeableness as additional covariates whenever a novel SNP-trait association shows an otherwise significant  $p$ -value.

## **Acknowledgments**

We thank Stephen Kosslyn for his support.

## **2 $g$ and the Psychological Refractory Period: Individual Differences in the Mind's Bottleneck**

James J. Lee<sup>1</sup>, Christopher F. Chabris<sup>2</sup>

**1 Department of Psychology, Harvard University, Cambridge, MA, USA**

**2 Department of Psychology, Union College, Schenectady, NY, USA**

### **Abstract**

Higher levels of general mental ability ( $g$ ) are associated with faster reaction times in elementary cognitive tasks. Here we pinpoint the locus of this association within a partition of reaction time into distinct processing stages. We adopted a number-comparison task permitting both experimental manipulation of multiple stages and the near-simultaneous presentation of two stimuli. Among the three stages distinguished by our experimental paradigms—perception, decision-making, and motor execution—it is only the central decision-making stage where higher- $g$  individuals enjoy an advantage. First, the only manipulation statistically interacting with  $g$  is already known to affect a central stage. Second, the advantage of the higher- $g$  individuals in responding to the second stimulus within the dual task doubled when the stimuli were presented very close together in time, which indicates that the advantage inheres in a stage that can perform a computation for only one stimulus at a time.

This seriality (“refractoriness”) of the central decisional stage is a well-established finding of dual-task studies. Finally, a decomposition of reaction time into a diffusion of evidentiary strength between decision boundaries and a low-variability residual stage revealed that the *g*-speed association is attributable to diffusion rate. This *g*-diffusion association converges with our replication of the finding that the diffusion process is encompassed by the central stage identified by interaction and dual-task analyses. Thus, our results unify three strands of research: the psychological refractory period, diffusion modeling, and individual differences in high-level mental abilities. An agenda for future investigations is to determine whether the association between *g* and the duration of the central bottleneck implies a causal role of this elementary architectural feature in language comprehension, quantitative reasoning, spatial-visualization, and other complex *g*-loaded abilities.

## **Author Summary**

Parsing a mental operation into stages, identifying the parallel or serial nature of each stage, and characterizing stage-specific processing mechanisms are important goals of the brain and behavioral sciences. Two distinct theoretical approaches have produced partitions of reaction time in elementary cognitive tasks. One has divided the total response into parallel perceptual and motor stages surrounding a serial decision-making stage. Another has divided the response into the “hitting time” of a stochastic evidence accumulation and a low-variability residual. These two partitions have recently been unified by showing that the serial bottleneck giving rise to the psychological refractory period encompasses the stochastic portion of the diffusion model. It has also been shown that general cognitive ability (*g*) is

associated only with diffusion rate and not residual time. Combining these findings, we can deduce that the faster reaction times of higher-*g* individuals reflect an advantage only in the serial bottleneck of central processing and not in the parallel peripheral stages. In this study we verify this theoretical deduction through multiple lines of converging evidence.

## 2.1 Introduction

Psychologists who study individual differences have been reasonably successful in quantifying variation in mental abilities with the use of standardized instruments that go under various names, including “IQ” or “scholastic aptitude” tests. Such a test is an aggregate of items eliciting responses that can be unambiguously scored as *right* or *wrong*. It is a remarkable fact that the responses to almost all such items, regardless of the specific skills or knowledge required, are positively correlated (Guttman & Levy, 1991). As a consequence a sample of items provides information about how the examinees would have performed on the much greater number of items that were not administered to them. This is one reason why an overall IQ score can be claimed to represent a valid measure of a person’s intelligence: under certain mathematical conditions that are reasonably well satisfied by actual mental tests, an examinee’s observed score on a test of increasing length approaches the score that he would have obtained on an infinitely long test covering all subdomains of logical, factual, and semantic knowledge (Mulaik & McDonald, 1978). In order to generalize beyond the score obtained on a particular test to mastery of this idealized wider domain, differential psychologists refer to the random variable mapping into the latter as “the *g* factor” (Spearman, 1927; Humphreys, 1994; Jensen, 1998).

*g* is correlated with a number of neural variables, including overall brain volume, con-

nectivity of white matter, and concentrations of *N*-acetyl aspartate (Chabris, 2007; Jung & Haier, 2007; Deary et al., 2010). However, before we can attempt to understand the causal structures underlying these correlations, it may be necessary to validate intermediate bridging laws expressed in terms of mental architecture and elementary cognitive operations (Sternberg, 1977; Hunt, 1978; Deary, 2001; Hunt, 2005; Conway et al., 2007). One strand of research has looked to the correlation between *g* and reaction time (RT) in simple laboratory tasks for clues to bridging laws of this kind (Jensen, 2006). It is now well established that higher levels of *g* are associated with faster mean times and lower variability across trials (Hunt et al., 1975; Vernon, 1983; Jensen, 1987a,b; Miller & Vernon, 1992; Deary et al., 2001). But until recently there have been relatively few attempts to locate the locus of this association within models partitioning the flow of information between perception and action into distinct processing stages (Luce, 1986; Sanders, 1998; Pashler, 1998). Because the distinctive properties of different stages are potentially revealing with respect to deeper mechanisms, the integration of the *g*-RT correlation into stage models is a promising avenue for the tracing of individual differences to lower-level causes.

There are two distinguishable types of RT partitions to which one might turn. The first turns on the distinction between parallel and serial processing raised by studies of so-called dual tasks, in which participants must respond to two stimuli presented close together in time. At very short delays, the RT to the second stimulus becomes longer (Telford, 1931; Welford, 1980; Pashler & Johnston, 1998; Lien et al., 2006). A parsimonious account of this *psychological refractory period* (PRP) invokes three successive stages of processing: a perceptual stage (*P*), a central stage (*C*), and a motor stage (*M*). The *P* stage consists of a translation of raw sensory input to a more abstract format that can be broadcast to down-



stream processors unconcerned with retinal locus, stimulus-background contrast, character font, and other such low-level features. The *C* stage consists of a mapping from percept to response, and PRP theory posits that this stage alone gives rise to the delay observed for the second RT in a dual task (Figure 3). The final *M* stage consists of implementing the motor response selected by the *C* stage. The corollary of the *C* stage being the only bottleneck is that, up to a certain limit, perceptual and motor processing for a given stimulus can take place concurrently with processing of any kind for another stimulus.

A second line of research has attempted to explain the characteristic dispersion and skewness of RT distributions. The most successful models of two-choice tasks posit a partition of RT into a stochastic process deciding between the two responses and a low-variability residual time (Ratcliff, 2002; Ratcliff & Smith, 2004; Ratcliff & McKoon, 2008). During the stochastic process, an internal variable undertakes a *diffusion* (continuous random walk) between two boundaries, each of which corresponds to a response alternative (Figure 3). The diffusion can thus be interpreted as the noisy accumulation of evidence in favor of one response. The accumulation terminates when it reaches one of the decision boundaries, which are set sufficiently far apart to ensure that absorption at the correct response occurs with high probability. Variants of this diffusion model have attracted much attention in recent years, not only because of their excellent fit to human behavioral data, but also because each parameter appears to have a distinct neural basis (Gold & Shadlen, 2007; Heekeren et al., 2008; Sajda et al., 2011). For example, studies of trained monkeys have recorded from neurons whose firing rates over the course of a RT trial trace a stochastic process predicting the monkey's motor response (Hanes & Schall, 1996; Gold & Shadlen, 2000; Ratcliff et al., 2007).

After decades during which these three research programs (individual differences, PRP, stochastic modeling) proceeded in parallel streams with scarcely any cross-currents, a series of seminal investigations have begun unifying them into a coherent whole. First, it has been shown that the stochastic evidence-accumulation stage of the diffusion model is encompassed wholly within the serial *C* stage of the PRP model (Sigman & Dehaene, 2005, 2006). Second, it has been shown that the *g*-RT correlation reflects a correlation of *g* with the rate of the diffusion process and not with residual time (Schmiedek et al., 2007; Ratcliff et al., 2008; Wagenmakers, 2009; Ratcliff et al., 2010; van Ravenzwaaij et al., 2011).

These results harmonize with many related proposals dividing human mental architecture into two broad components: (1) a number of parallel (modular) processors of sensory data and motor commands, operating inflexibly but with great precision; and (2) a central workspace that can establish arbitrary links between processors through a slow, variable, and serial chain of computations (Fodor, 1983; Shallice, 1988; Dennett, 1991; Baars, 1997; O'Reilly, 2006; Baddeley, 2007; Dehaene, 2008). Further consideration of the resulting unified model leads to a clear deduction regarding the properties of the *g*-RT correlation that has so far been untested: the serial bottleneck posed by *C* is the only stage in the PRP partition that contributes to the correlation. In fact, this is the deduction to be tested in the present work. While the hypothesis that *g* is correlated only with *C* and not with *P* or *M* is simple to state, it generates a number of stringent predictions for our experiments manipulating the demands on specific stages and varying the time between stimulus onsets within a dual task.

We adopted a particular RT task used in many previous studies of numerical cognition or dual-task performance (Moyer & Landauer, 1967; Dehaene, 1996; Logan & Gordon, 2001; Sigman & Dehaene, 2005, 2006, 2008; Corallo et al., 2008; Song & Nakayama, 2009; Hes-

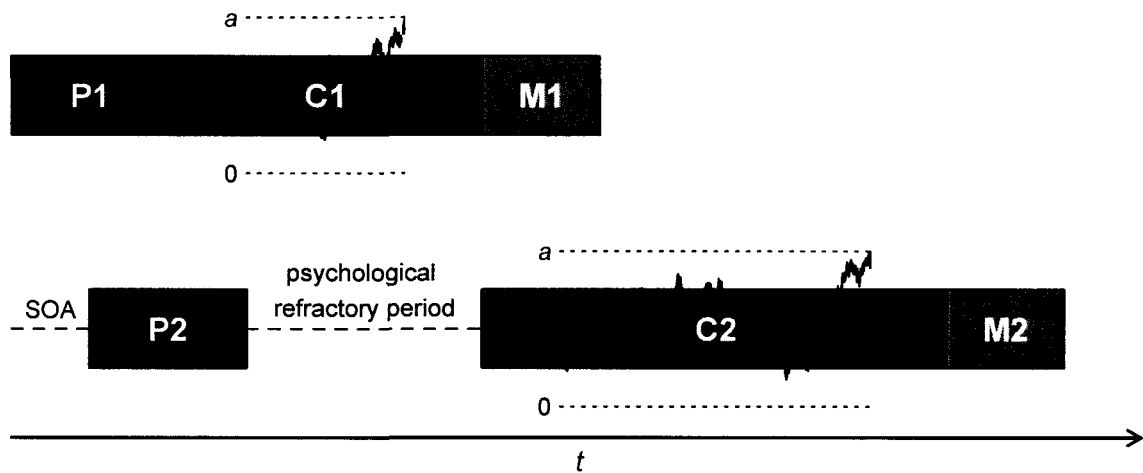


Figure 3: **A unified model of the psychological refractory period (PRP) and diffusion of evidentiary strength in two-choice RT tasks.** The second stimulus follows the first after a brief stimulus onset asynchrony (SOA). The perceptual (*P*) and motor (*M*) stages vary little in duration from trial to trial and can be carried out in parallel with stages of another task. The central (*C*) stage contains a noisy accumulation of evidence (diffusion) until a decision threshold is reached. *C*<sub>2</sub> cannot start until *C*<sub>1</sub> is finished, which results in the bottleneck (“refractoriness”) referred to as the PRP. Increasing the distance between the decision thresholds (*a*) results in greater accuracy but a longer processing duration. Although the depicted diffusions were simulated with the same rate, the diffusion in *C*<sub>2</sub> took nearly twice as long to reach absorption. This shows the intrinsic variability of central processing. If *g* is positively correlated with diffusion rate, it follows that *g* is also associated with more rapid and reliable progress through the serial bottleneck.

selmann et al., 2011). The task requires deciding, quickly and accurately, whether a number presented on a computer screen is smaller or larger than a reference number. There now exists a broad body of mathematical theory and neuroscientific evidence regarding the mental representation and processing of number (Gallistel & Gelman, 2005; Dehaene, 2007), which makes this task a promising tool for advancing a mechanistic account of individual differences. We administered different versions of the number-comparison task to volunteers who qualified for the study by reporting either a score of 1560 or higher on the combined verbal and mathematics sections of the SAT (formerly Scholastic Aptitude Test) or a score of 1280 or lower. Although lacking a component testing spatial ability, the SAT is otherwise an excellent measure of  $g$  (Frey & Detterman, 2004). We refer to the participants reporting a score of 1560 or higher as the “high- $g$ ” group and to the other participants as the “moderate- $g$ ” group. The cut score of 1280 corresponds to approximately the 89th percentile of SAT scores, and the difference between the high and moderate cut scores is about 1.3 standard deviations with reference to the total population of examinees taking the SAT.

## **2.2 Results**

In some cases we were not able to verify a participant’s self-reported SAT score with the university registrar or a College Board report. Therefore we validated the self-reports with administrations of brief IQ tests during the study session. Although these in-laboratory tests contain fewer items than the SAT and hence are less reliable, they still almost perfectly separated the moderate- and high- $g$  groups (Figure 4).

We organize the remainder of the Results as follows. We first analyze the means and variances of number-comparison RT under various experimental manipulations. By examin-

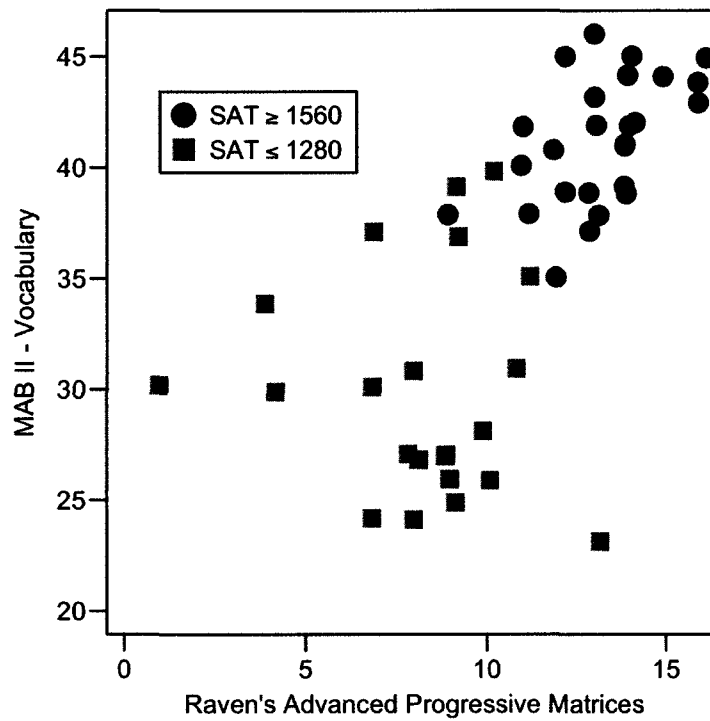


Figure 4: **Validation of self-reported SAT scores as a measure of  $g$ .** Raven's Advanced Progressive Matrices and the Multidimensional Aptitude Battery II are both IQ tests in widespread use. Plotted are the jittered number of items answered correctly.

ing the pattern of statistical interactions, we can discover the number of stages distinguished by the manipulations and which of these accounts for the  $g$ -RT correlation. A diffusion decomposition of RT supplies an independent check of the interaction analysis. If the pattern of statistical interactions points to a particular stage as responsible for the  $g$ -RT correlation, then  $g$  should be associated with only the diffusion parameters governing the duration of this stage. Next, by determining the nature of the interference between two stimuli within a dual version of the number-comparison task, we verify the temporal ordering of the major stages and whether each stage can be executed in parallel with the processing of the other stimulus. Depending on whether the  $g$ -associated stage is serial or parallel, we arrive at different point predictions of the difference between the moderate- and high- $g$  groups as a function of the elapsed time between stimulus onsets. A comparison of these predictions with our data shows that the  $g$  difference resides in the same central decisional stage identified by the interaction and diffusion analysis.

For reference we provide definitions of all symbols and terminology in Table 7.

### **2.2.1 Analysis of Single-Task RT Means and Variances**

Suppose that the overall RT in the number-comparison task is indeed the sum of three serially arranged stages with constant outputs (Figure 5A). By “constant” we mean that it is possible, loosely speaking, to label stage 1’s outputs in such a way that the processing within stage 2 only depends on the label and not on the particulars of the path taken through stage 1. For example, in the context of the number-comparison task, we say that the output of  $P$  is constant if  $C$  is affected only by the numerical magnitude of the stimulus identified by  $P$  and not by other stimulus features such as color or font. A critical property of serially arranged

Table 7: Symbols and terminology.

---

$g$	“General intelligence” or “general mental ability.” A person’s level of $g$ is the score, expressed on a convenient metric, that he would have obtained on an infinite number of items testing language comprehension, quantitative reasoning, reasoning with abstract patterns, spatial visualization, and so on.
SAT	Formerly, the Scholastic Aptitude Test. A test of verbal and mathematical ability used in the US to screen college applicants. A good measure of $g$ .
RT1, RT2	The respective reaction times to the first and second stimuli within a dual task. Each RT is measured from the onset of its own stimulus.
SOA	Stimulus onset asynchrony. The experimentally controlled temporal interval between the onsets of the two stimuli in a dual task.
PRP	Psychological refractory period. The prolonging of RT2 observed when the SOA is very short.
$P, C, M$	Respectively the perceptual, central, and motor stages of the RT to a given stimulus. Subscripts are used if necessary to distinguish between the first and second stimuli in a dual task.
additive factors	A methodology used to infer the existence of separate processing stages underlying RT. If the response process is composed of serially arranged stages with constant outputs, two manipulations affecting different stages should be additive in their effects. If two manipulations affect the same stage, they should show a statistical interaction.
diffusion	The movement over time of some variable that is subject to many small perturbations; the limit of a random walk as the steps in both space and time become extremely small. In the present context, the noisy accumulation of evidence in favor of one response alternative over the other.
$T, a, v$	The parameters of RT diffusion modeling. $T$ is the duration of the low-variability residual stage (according to the unified model in Figure 3, the summed durations of $P$ , $M$ , and the non-diffusion portions of $C$ ). $a$ is the separation between the decision boundaries. $v$ is the rate at which the process would travel from the starting point ( $a/2$ ) to the appropriate boundary in the absence of stochastic perturbations.

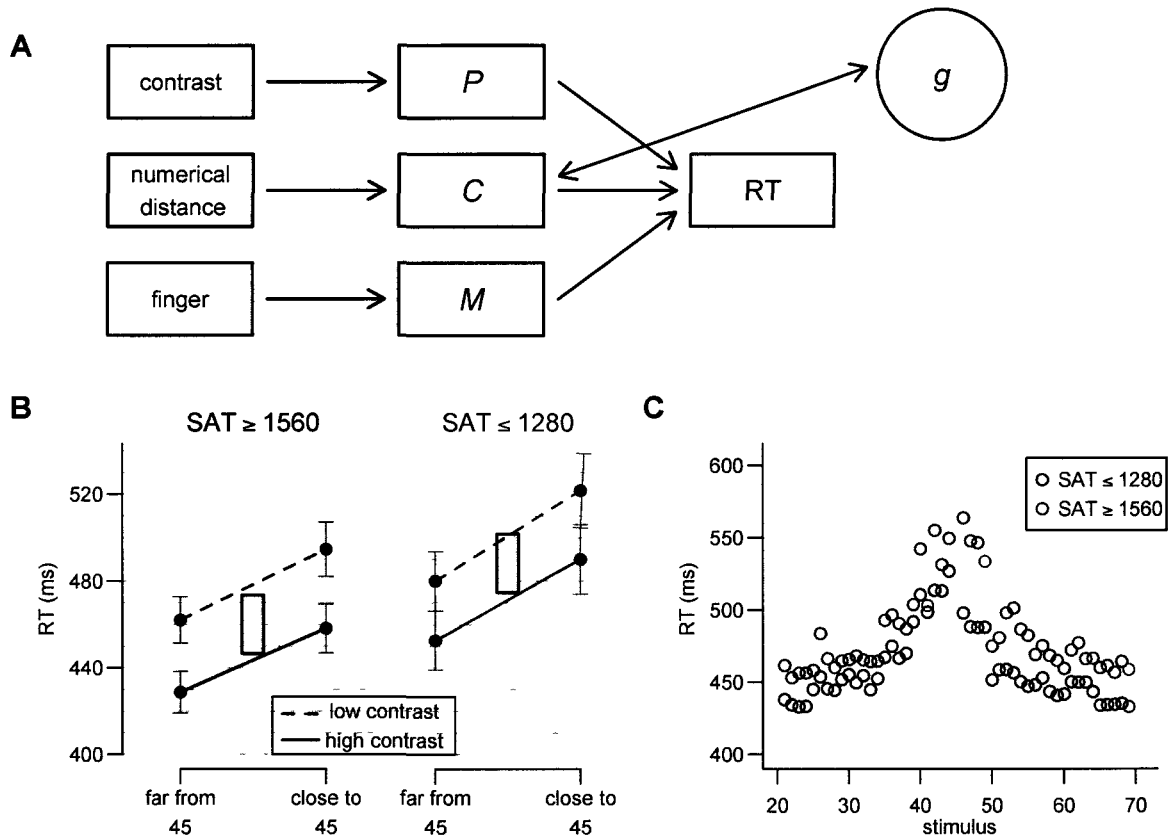
---

stages with constant outputs is that we can regard the stages as wholly distinct variables, each of which can be shortened or prolonged independently of the others (Sternberg, 1969; Sanders, 1990; Roberts & Sternberg, 1993). Therefore experimental manipulations of distinct stages cannot show any deviations from additivity. On the other hand, manipulations affecting the same stage will in general show a non-additive interaction. We should also observe an interaction with measurements of naturally occurring variation if a manipulation affects a stage associated with this variation.

We used this logic of additive factors to ascribe naturally occurring variation in  $g$  to a distinct processing stage. In our first experiment, participants mapped the numbers presented on a computer screen that were less than 45 in magnitude to a key press with their left hands and the numbers larger than 45 to a key press with their right hands. Within blocks the numbers were either black or light gray, thus contrasting either sharply or hardly at all with the white background. We expected this manipulation of stimulus-background contrast to affect the  $P$  stage. The numbers ranged in magnitude from 21 to 69, excluding 45, and we expected this manipulation of numerical distance to affect the  $C$  stage. At the beginning of each block, participants were instructed to respond with either their index fingers only or their ring fingers only, and we expected this manipulation to affect the  $M$  stage. An additional postulate is that individual differences in brain structure or function affecting the  $C$  stage of this artificial task also affect central processing in natural settings, leading to the high-level individual differences summarized as  $g$ . Combining these hypotheses as in Figure 5A, we arrive at the following prediction for the outcome of the experiment: the only non-additive interaction among  $g$ , contrast, distance, and finger should be that between  $g$  and distance.

The results of the experiment are summarized in Table 8 and Figure 5B. Contrary to our





**Figure 5: The use of additive factors to assign the  $g$ -RT correlation to a stage affected by a particular manipulation.** (A) A directed acyclic graph representing the causal relations among the observed and hypothesized variables in the additive-factors experiment. (B) Cell means across the experimental manipulations and levels of  $g$ . For clarity the manipulation of the response finger has been omitted. The height of the gold bar corresponds to the estimated effect of the contrast manipulation (27 ms). For this plot numerical distance was dichotomized into “far” ( $|\text{stimulus} - 45| > 12$ ) and “close” ( $|\text{stimulus} - 45| \leq 12$ ). The bars encompass  $\pm 1$  standard error. (C) Mean RT as a function of numerical magnitude.

**Table 8: Coefficients of  $g$  and manipulations in linear mixed models of RT means and variances.**

predictor	mean (ms)	variance (ms <sup>2</sup> )
low contrast ( $P$ )	<b>27.05 ± 4.38</b>	1355 ± 1349
numerical distance ( $C$ )	<b>-3.49 ± 0.21</b>	<b>-6734 ± 1349</b>
ring finger ( $M$ )	-5.48 ± 5.08	-153 ± 1349
high $g$	<b>-40.75 ± 18.5</b>	<b>-4997 ± 2183</b>
$P \times C$	-0.14 ± 0.30	-1277 ± 1908
$P \times M$	-2.09 ± 6.22	149 ± 1908
$C \times M$	-0.11 ± 0.30	560 ± 1908
$P \times g$	7.22 ± 5.75	-573 ± 1784
$C \times g$	<b>0.88 ± 0.28</b>	<b>3727 ± 1784</b>
$M \times g$	8.85 ± 6.68	851 ± 1784
$P \times C \times M$	0.51 ± 0.43	104 ± 2698
$P \times C \times g$	-0.05 ± 0.40	-48 ± 2523
$P \times M \times g$	-2.44 ± 8.16	-1119 ± 2523
$C \times M \times g$	-0.34 ± 0.40	-1412 ± 2523
$P \times C \times M \times g$	-0.14 ± 0.56	1118 ± 3569

For brevity each manipulation is symbolized by the stage that it putatively affects (Figure 5A). The model for the means was fit to the individual trial data. To fit the model to the cell variances, we dichotomized numerical distance. Each estimated coefficient is given with plus/minus its standard error.

expectations from a pilot study, the finger manipulation proved quite heterogeneous. Analyzing each individual separately, we found statistically significant effects of this manipulation in opposite directions. We thus treated both the intercept and the deviation from the mean effect of finger use as random effects in a linear mixed model. Because the small-sample distribution of the fixed effects is not known, we cannot give exact model-based  $p$ -values. Heuristically we regard any coefficient exceeding twice its standard error in absolute value as statistically significant.

The findings regarding mean RT conform to our predictions. (1) Contrast and numer-

ical distance had significant main effects. Reducing the stimulus-background contrast and decreasing the absolute stimulus-reference numerical distance both slowed RT. Treating numerical distance as a dichotomous variable, we see that these two manipulations showed comparable effect sizes (Figure 5B). (2) All pairwise and higher-order interactions among manipulations were nonsignificant. Furthermore, we could not reject a model setting all nonsignificant interactions to zero in favor of a model with all interactions free ( $\chi^2_{10} = 13.6$ ,  $p > .18$ ; both AIC and BIC smaller for the more parsimonious model). The data are thus compatible with our postulate that the number-comparison task is composed of three ordered and separately manipulable stages. (3) As expected, the high- $g$  participants responded more rapidly. They also made fewer errors (accuracy .943 vs. .929), indicating that speed-accuracy tradeoff alone is unlikely to account for their advantage. (4) The only significant interaction involving  $g$  was the pairwise interaction between  $g$  and numerical distance. Figure 5C reveals the nature of this interaction. As the numerical magnitude of the stimulus approached the reference, all participants tended to slow down. This trend was more pronounced in the moderate- $g$  participants, however, leading to a greater advantage of the high- $g$  group for stimuli close to the reference. This selectivity of  $g$ 's behavior with respect to the experimental manipulations is consistent with our hypothesis that the  $g$  difference in RT is a difference in the  $C$  stage alone.

If the stage durations are stochastically independent, then the logic of additive factors applies to the variances as well as the means. Table 8 shows that the variances also conform to our model: the only manipulation interacting with  $g$  was numerical distance. We could not reject a model setting all nonsignificant interactions to zero in favor of a model with all interactions free ( $\chi^2_{10} = 3.09$ ,  $p > .97$ ; both AIC and BIC smaller for the more parsimonious

model). The nature of the *g*-distance interaction mirrors that for the means; all participants became more variable as the stimulus approached the reference, but this trend was stronger in the moderate-*g* participants. Neither contrast nor finger significantly affected RT variance. Altogether these results suggest that whereas changes in perceptual features and the response effector alter intrinsically low-variability stages, both *g* and changes in semantic content are associated with a stochastic process where the variance increases along with the mean (Hunt et al., 1975).

2 1 2 Diff. in RT of Stim. to Ref. RT

Table 9: Associations of  $g$  and manipulations with diffusion parameters.

predictor	$T$	$a$	$v$
contrast	<b>27.5 ms</b> $p < 2 \times 10^{-9}$ <b>95% CI = (22.9, 33.2)</b>		$-.015 s^{-1}$ $p > .29$ 95% CI = (-.044, .011)
distance	5.5 ms $p < .005$ 95% CI = (1.6, 9.5)		<b>.193 <math>s^{-1}</math></b> $p < 5 \times 10^{-8}$ <b>95% CI = (.159, .250)</b>
finger	-4.3 ms $p > .16$ 95% CI = (-9.7, 2.0)	$\sim 0$ $p > .93$ 95% CI = (-.004, .004)	$-.037 s^{-1}$ $p < .02$ 95% CI = (-.104, -.010)
$g$	10 ms $p > .35$ 95% CI = (-11.2, 31.1)	$-.012$ $p > .13$ 95% CI = (-.030, .003)	<b>.115 <math>s^{-1}</math></b> $p < .001$ <b>95% CI = (.038, .286)</b>

Numerical distance was trichotomized into “close” ( $|\text{stimulus} - 45| \leq 8$ ), “intermediate” ( $8 < |\text{stimulus} - 45| \leq 16$ ), and “far” ( $|\text{stimulus} - 45| > 16$ ) in order to meet the recommended 125 trials per cell (Grasman et al., 2009). An effect of distance should thus be interpreted as the average increment per ordered level. We used high contrast, far distance, index finger, and low  $g$  as the respective baselines for assessing differences. Since the average effect of ring-finger use on  $a$  was virtually zero, we used the average  $a$  over blocks to examine whether the  $g$  groups differed in speed-accuracy tradeoff.  $g$  did not significantly interact with any manipulations.

effect only on  $T$ . Moreover, the estimated average effect over participants of 27.5 ms is remarkably close to the estimated effect on total RT in the additive-factors analysis. (2) The average effect of finger on  $T$  is similar to the estimated effect on total RT in the additive-factors analysis. Also consistent with our analysis of mean RT is the heterogeneity of this effect across individuals; the estimates range from  $-31$  to  $30$  ms with a suggestion of bimodality. (3) Both numerical distance and  $g$  showed statistically and quantitatively significant associations with  $v$ . (4) Any difference in  $T$  between the moderate- and high- $g$  groups made no detectable contribution to the overall  $g$ -RT association.

Overall, these results suggest that  $g$  and numerical distance are associated with a decisional stage depending on a noisy accumulation of evidence, whereas contrast and finger affect low-variability stages whose summed durations are less than or equal to  $T$ . This diffusion decomposition parses RT in exactly the same way as the additive-factors analysis, thus confirming and extending our earlier results.

We now discuss two anomalies in Table 9. First, response finger showed a small but significant effect on  $v$ . We address this deviation from our model in the Discussion. Second, numerical distance showed a significant effect on  $T$ . Examination of Figure 5C suggests the reason for this effect: both moderate- and high- $g$  groups showed discontinuities in mean RT when the stimulus crossed the edges of the decade including the reference. To confirm that the effect of distance on  $T$  reflects a perceptual cost of distinguishing the stimulus from the reference when both begin with the same tens digit (Dehaene, 2007), we estimated the diffusion parameters for each participant after rebinning the stimuli into five ordered levels coinciding with decade borders and midway points. In this analysis we collapsed the data across levels of the contrast and finger manipulations. The results are plotted in Figure 6.

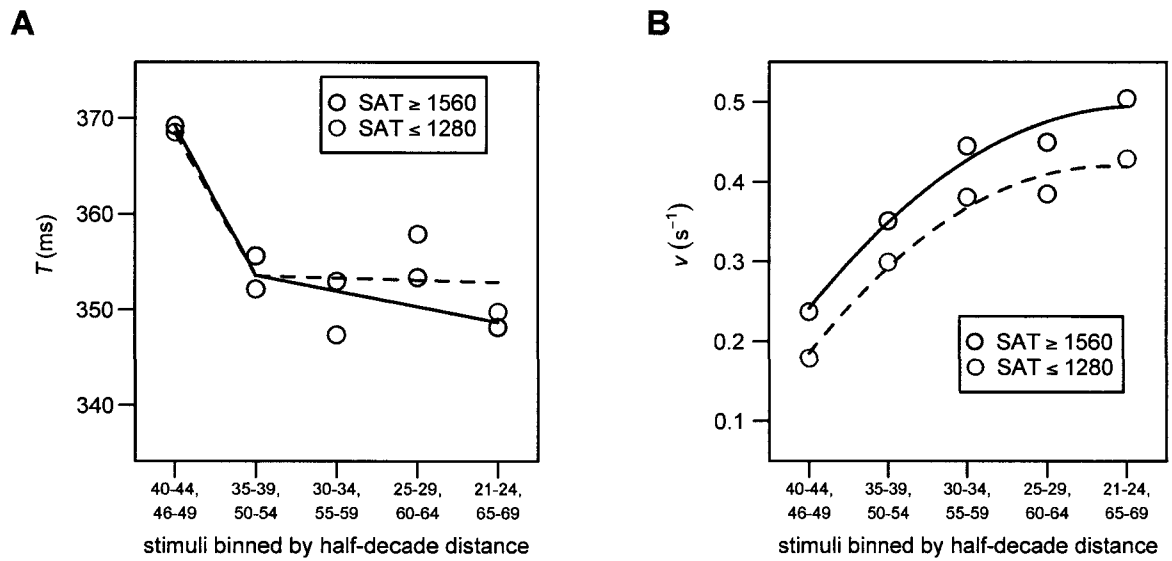


Figure 6: **Estimates of mean diffusion parameters as a function of binned numerical distance.** (A) Mean values of  $T$  (fixed residual time). The spike when the stimulus crosses the edges of the decade including the reference suggests an added perceptual cost of the tens digits being identical in appearance. This cost was virtually the same in the moderate- and high- $g$  groups. (B) Mean values of  $v$  (diffusion rate). Throughout the full range of numerical distance, the rates followed smooth curves displaced from each other by a constant in favor of the high- $g$  group.

Whereas the mean values of  $T$  showed no trend outside of the 40s ( $\beta_{\text{distance}} = -.3 \pm 1$  ms/bin), there were clear discontinuous steps at the edges of this decade [16 ms;  $p < 7 \times 10^{-4}$ ; 95% CI = (9.5, 23.7)]. In contrast, there were no discernible discontinuities in the mean values of  $\nu$ . The regression lines in Figure 6B include quadratic terms; we will later discuss the implications of this curvature. For now it suffices to note that a linear increase in  $\nu$  with distance continued beyond the edges of the 40s and throughout the 30s and 50s. This difference between  $T$  and  $\nu$  in notational sensitivity attests to the strict modularity of perceptual and central processing. Unaffected by the knowledge that the ones digit of the stimulus is irrelevant if the tens digit is not 4, the  $P$  stage always delivers to the  $C$  stage a full representation of numerical magnitude. The additional analysis devoted by  $P$  to recovering this magnitude if stimulus and reference share the same tens digit does not benefit the  $C$  stage, which processes number without regard for the historical contingencies that have left some pairs of number symbols much more perceptually similar than others. In short,  $P$  cannot “know” whereas  $C$  cannot “see.”

It is evident that  $g$  does not interact with the tens digit. Thus, instead of contradicting our hypothesis that the  $g$ -RT correlation resides in one of many separable processing stages, the effect of distance on  $T$  turns out to provide unexpected support for it.

### 2.2.3 Analysis of Dual-Task Means, Variances, and Correlations

Without relying on prior findings, our results so far would support the following: (1) our experimental manipulations affect three or four distinct processing stages, (2) the difference between the  $g$  groups in RT is a difference in the one stage affected by notation-invariant numerical distance, and (3) this latter stage consists of a stochastic process where the vari-



ance increases along with the mean. We would not be able to infer, however, whether the  $g$ -associated stage can proceed in parallel with the processing of another stimulus presented closely in time. Even our supposition regarding temporal order is based as much on logical considerations as on empirical ones.

To ascertain the temporal and time-sharing properties of the  $g$ -associated stage, we performed a second experiment, presenting the number-comparison task as a dual task to an independent sample of participants. The numbers ranged in magnitude from 1 to 9, excluding 5, and participants had to judge whether the presented number was smaller or larger than 5. One number appeared just to the left of the fixation cross, and participants had to map *smaller* to a key press with the left middle finger and *larger* to a key press with the left index finger. After an SOA of 60 to 960 ms following the onset of the first number, a second number appeared just to the right of the fixation cross. With respect to this second number, participants had to map *smaller* to a key press with the right index finger and *larger* to a key press with the right middle finger. Thus, each hand corresponded to the numbers appearing on its side of the fixation cross, and within each hand *smaller* mapped to the leftmost finger and *larger* to the rightmost. Participants were instructed to respond to each number as quickly as possible while still being highly accurate.

Before presenting the primary results of this dual-task experiment, we give its asymptotic EZ2 estimates to verify our diffusion analysis of single-task RT (Table 10). The dual-task data replicated the additive-factors data in two key respects: (1)  $g$  was scarcely associated with  $T$ , and (2)  $g$  positively covaried with  $\nu$ . As required by the latter finding, with respect to total RT there was a significant interaction of  $g$  and numerical distance, despite the restricted range of variation in the latter variable ( $\beta_g = -81 \pm 23$  ms;  $\beta_{\text{distance}} = -22 \pm$

Table 10: Diffusion parameters of asymptotic RT2 (SOA  $\geq$  900 ms).

parameter	individual estimates														
<i>T</i> (ms)															
high <i>g</i>	268	283	293	313	321	324	326	328	330	332	334	334	350	365	
low <i>g</i>	240	289	294	309	311	314	332	342	346	349	350	354	369	377	
	-5, $p > .67$ , 95% CI = (-26, 19)														
<i>a</i>															
high <i>g</i>	.065	.067	.075	.077	.079	.080	.088	.088	.090	.093	.093	.098	.108	.119	
low <i>g</i>	.076	.079	.080	.081	.086	.088	.096	.102	.113	.121	.126	.134	.137	.203	
	-.022, $p < .004$ , 95% CI = (-.046, -.006)														
<i>v</i> ( $s^{-1}$ )															
high <i>g</i>	.29	.31	.34	.34	.36	.36	.36	.37	.40	.40	.45	.45	.50	.54	
low <i>g</i>	.23	.26	.30	.31	.31	.33	.36	.37	.37	.39	.39	.39	.41	.43	
	<b>.045, <math>p &lt; .04</math>, 95% CI = (.003, .095)</b>														

The difference in means between the moderate- and high-*g* groups, along with its *p*-value and 95% confidence interval, are given below the estimates of each parameter. We grouped together all trials at SOA  $\geq$  900, collapsing over numerical magnitude, to meet the recommended 125 trials.

2 ms;  $\beta_{g \times SOA} = 9.3 \pm 3.0$  ms). Another noteworthy finding is that our high-*g* participants in the dual-task experiment were more reckless, showing a significantly smaller mean value of *a*. Different studies have now found *g*-*a* associations of opposite sign (Schmiedek et al., 2007; Wagenmakers, 2009; Ratcliff et al., 2010), suggesting that this relationship depends on the instructions given in a particular experiment.

If all processing stages are parallel, then RT1 and RT2 should have identical distributions. In contrast, if all processing stages for the second stimulus must await the response to the first stimulus, then simultaneous presentation should lead to RT2 being twice as long as RT1. As a matter of fact, dual-task experiments almost always find that a very short stimulus onset asynchrony (SOA) leads to RT2 being longer than RT1 but by less than a factor of two.

This observation implies that both parallel and serial stages make a contribution to overall RT.

Figure 7 shows how the precise nature of dual-task interference reveals the relative position of each stage and whether it can be parallelized (Schweickert & Townsend, 1989; Pashler, 1994). The insets are theoretically predicted graphs of RT1 and RT2 as a function of SOA, at both “fast” and “slow” levels of the relevant variable. Given the assumption of a single serial bottleneck, a parallel stage must either precede the bottleneck or follow it. Panels A and C show that any elongation of a stage up to and including the bottleneck will increase both RT1 and RT2 at a short SOA. Once the SOA is long enough for the two serial portions to be well clear of each other, this effect of manipulating the first stimulus on RT2 vanishes. A key point is that changes in RT1 and RT2 upon a manipulation of the first stimulus cannot reveal whether the manipulation affects a serial stage or a parallel stage preceding it. The propagation into RT2 occurs because its serial portion can begin no sooner than the termination of RT1’s serial portion, which can be pushed forward by an elongation of either a preceding stage or the serial portion itself.

Manipulating the second stimulus leads to dramatically different RT2 profiles, however, depending on whether the manipulation prolongs a pre-bottleneck stage or the bottleneck itself. As might be expected, prolonging the serial portion of RT2 has no effect on RT1 and increases RT2, regardless of SOA (Figure 7D). In contrast, at a sufficiently short SOA, prolonging a pre-bottleneck stage of RT2 is predicted to have no effect on overall RT2 whatsoever (Figure 7B). This is because a postponement of RT2’s serial portion leaves a “slack” into which the duration of a pre-bottleneck stage can be expanded without pushing forward any subsequent stages.

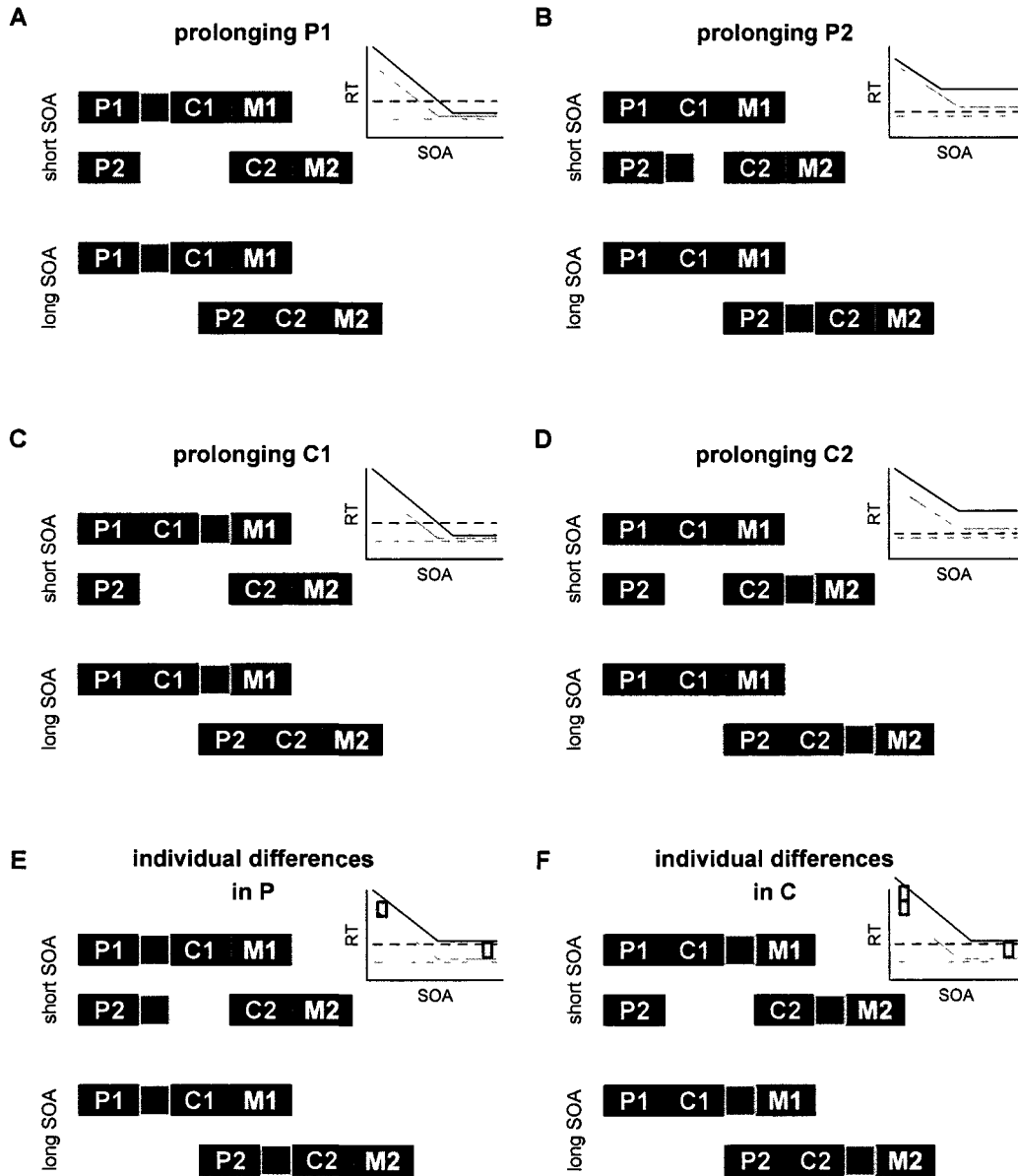


Figure 7: **The predictions of the PRP model.** Dashed and solid lines in the insets represent theoretical predictions of RT1 and RT2 respectively. Black and gray lines represent the “slow” and “fast” levels of the relevant variable. See the text for detailed explanations.

For brevity we have not depicted the predictions that follow from a manipulation of a post-bottleneck stage. But it is easy to verify that such a manipulation affects only the RT to the manipulated stimulus regardless of SOA. Thus, by performing dual-task experiments manipulating the demands imposed by both the first and second tasks, the resulting profiles of RT1 and RT2 as a function of SOA constitute a distinctive signature of the affected stage's position in the information flow. That is, the PRP model assigns each possible pattern of interference to a temporal position preceding, within, or following the serial contribution to RT.

PRP studies have typically manipulated stages of only one task at a time. Individual differences associated with a given RT stage, however, are analogous to a simultaneous manipulation of both tasks. Panels E and F show that this situation also leads to distinctive predictions. For simplicity we treat the case of the two tasks being the same. If individuals differ in the duration of a pre-bottleneck stage, then the difference is propagated only once into RT2 as a result of expansion into slack (Figure 7E). Therefore the difference between individuals in RT2 will not depend on SOA. One can easily see that the same invariance holds for a difference in a post-bottleneck stage or indeed for a combination of differences in any parallel stages surrounding the bottleneck.

If individuals differ in the duration of a serial stage, then at a short SOA this difference is propagated twice into RT2. In this temporal regime, the two serial computations are arranged end-to-end, and therefore the slower individual's RT2 suffers a double cost. This leads to a clear prediction: any difference between individuals in a serial stage will result in an exact doubling of their RT2 difference as the SOA diminishes to small values (Figure 7F).

If individuals differ in both parallel and serial stages, it is possible to observe other fac-

tors besides unity or two by which their difference in RT2 increases as the SOA diminishes. In these cases it is difficult to formulate sharp point predictions. However, given the now repeatedly replicated finding that  $g$  is associated with diffusion rate and not  $T$ , a division of the  $g$  difference across stages with different time-sharing properties is rather implausible. Such a division would amount to the diffusion of evidentiary strength between decision boundaries switching from seriality to parallelism while still in progress. A restriction of the  $g$  difference to the diffusion process is reinforced by the failure to detect associations between  $g$  and the components of  $P$  and  $M$  affected by our experimental manipulations. Thus, we have set up a confrontation between the point predictions of unity and two, and evidence in favor of the latter would support our deduction that  $g$  is associated with the rapidity of a serial processor.

Previous studies have shown that stimulus-background contrast affects a pre-bottleneck stage (Pashler, 1984; Pashler & Johnston, 1989; De Jong, 1993), numerical distance affects a serial stage (Sigman & Dehaene, 2005, 2006; Corallo et al., 2008), and motor demands affect post-bottleneck stages (Pashler, 1998; Ferreira & Pashler, 2002; Sigman & Dehaene, 2005). These findings justify the labeling of the ordered stages in Figure 7 as  $P$ ,  $C$ , and  $M$ . One issue that arises in our own study is whether the two stimuli can be perceptually processed in parallel despite being in the same modality (vision). Most recent PRP studies have used tasks impinging on different modalities (vision and audition). The theory of visual attention does suggest, however, that two spatially proximate alphanumeric targets on a blank computer screen should be fully processed in parallel (Pashler, 1998; Thornton & Gilden, 2007). To verify the parallelism of visual identification in our experiment, we administered a version of the number-comparison dual task varying the contrasts of both stimuli. Reducing

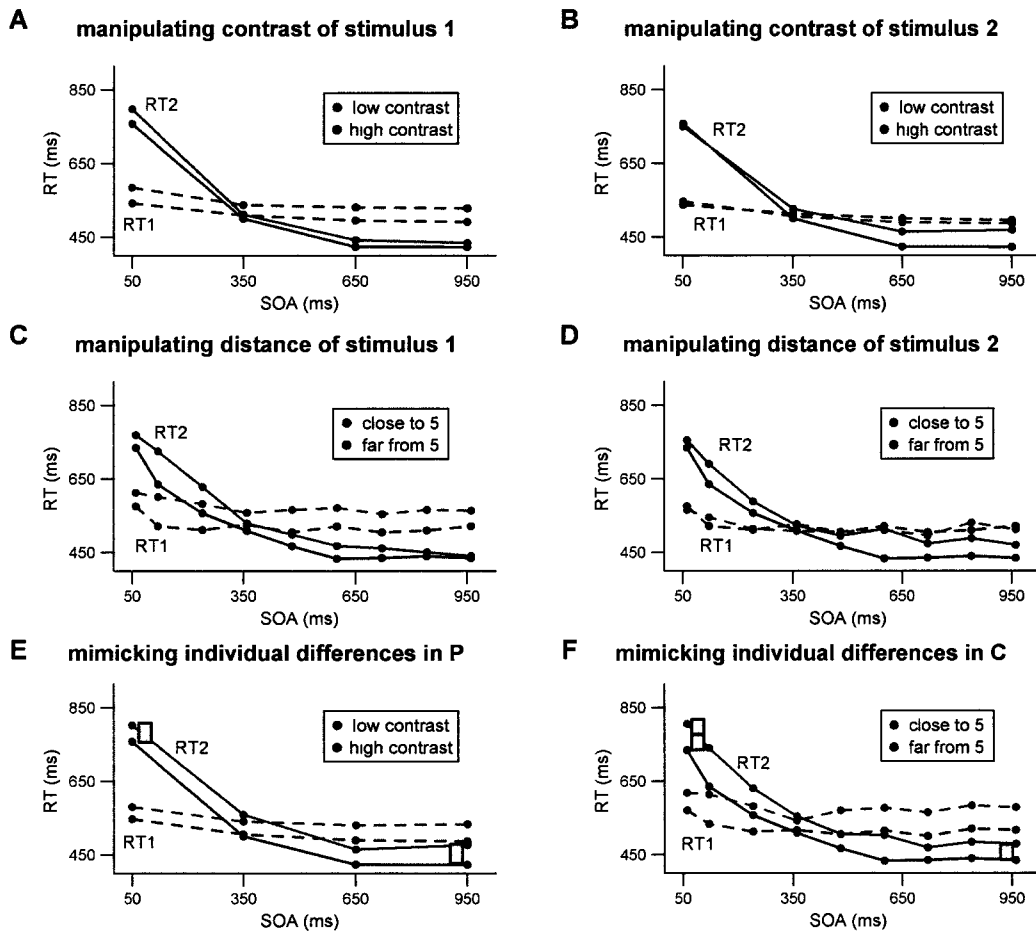


Figure 8: **Determining stage properties through dual-task interference.** For clarity panels C, D, and F display the data for only 9 of the 16 SOAs used. The height of the gold bars in E and F correspond to the effect of manipulating both stimuli at a long SOA ( $\geq 900$  ms). Compare with the theoretical curves in the insets of Figure 7.

Table 11: **Effects of manipulations on dual-task RT.**

	RT1 (ms)		RT2 (ms)	
	short SOA	long SOA	short SOA	long SOA
manipulating stimulus 1				
contrast	<b>42.4 ± 15.2</b>	<b>36.6 ± 12.0</b>	<b>39.3 ± 16.3</b>	11.3 ± 9.0
distance	<b>58.7 ± 7.5</b>	<b>45.6 ± 6.9</b>	<b>62.2 ± 8.3</b>	4.6 ± 3.6
manipulating stimulus 2				
contrast	-9.8 ± 13.9	9.4 ± 12.2	-7.6 ± 15.4	<b>45.5 ± 8.7</b>
distance	7.1 ± 6.5	-2.5 ± 5.9	<b>38.2 ± 8.0</b>	<b>42.2 ± 4.8</b>
manipulating both stimuli				
contrast	<b>32.9 ± 13.9</b>	<b>45.9 ± 13.1</b>	<b>44.4 ± 16.5</b>	<b>53.1 ± 9.4</b>
distance	<b>64.1 ± 8.1</b>	<b>67.0 ± 6.9</b>	<b>88.0 ± 9.5</b>	<b>43.0 ± 6.9</b>

Short and long SOA refer to  $\leq 120$  ms and  $\geq 900$  ms respectively. When assessing the effect of manipulating only one stimulus within a trial, the other stimulus was set at the “fast” level. Each estimate is given with plus/minus its standard error. Compare with the theoretical predictions in Figure 7.

the contrast of stimulus 1 delayed RT1 and also RT2 at short SOAs, which points to the *P* stage preceding or contributing to the serial bottleneck (Figure 8A). Reducing the contrast of stimulus 2 showed no effect on RT2 at short SOAs, however, confirming that *P* is indeed an early parallel stage (Figure 8B). Furthermore, in the version of the dual-task experiment presenting all stimuli at high contrast, we found that the same manipulation selectively affecting diffusion rate (numerical distance) also behaved as a determinant of a serial stage (Figures 8C and 8D). This is an important replication of the key finding that unifies the diffusion and PRP models (Sigman & Dehaene, 2005, 2006). Table 11 provides these results in numerical form.

An unexpected observation is that the distance manipulation exerted a consistently smaller effect when applied to the second stimulus. No prior results on stimulus crosstalk



seem to predict this particular interaction of distance and task order (Hommel, 1998; Logan & Gordon, 2001; Hesselmann et al., 2011), and we merely point out this anomaly without attempting to explain it.

Because our experiments contained trials with both stimuli set at the same difficulty level, we were able to mimic individual differences in the stages affected by our two manipulations. Whereas the effect on RT2 of a simultaneous reduction in contrast did not depend on SOA, the effect of a simultaneous reduction in numerical distance became much larger as the SOA diminished (Figures 8E and 8F; Table 11). These results are consistent with the parallelism of *P* and the seriality of *C*. They also presage our primary finding regarding the nature of the *g*-RT correlation.

Consider our main results to this point: (1) the dependence of the distance effect on *g*, (2) the association of both distance and *g* with diffusion rate, (3) the absence of an association between *g* and other processing stages, and (4) the seriality of the stage affected by distance. From these considerations it now follows that the difference between the moderate- and high-*g* groups resides in a serial stage and therefore must show up as a doubling of their difference in RT2 as the SOA becomes small. Figure 9 shows that in our sample the difference in RT2 associated with *g*, from SOA  $\geq 900$  to  $\leq 120$  ms, increased by a factor of 2.40 [99% CI = (1.60, 5.64)]. The interpretation of this factor, however, depends on the evident slowing of RT1 relative to asymptotic RT2 (Figures 8 and 9).

The commonly observed slowing of RT1 relative to single-task RT or asymptotic RT2 has been attributed to an executive task-scheduling stage (*E*) between *P*<sub>1</sub> and *C*<sub>1</sub> that decides the order in which to perform the central processing for the two stimuli (Meyer & Kieras, 1997; Jiang et al., 2004; Sigman & Dehaene, 2005, 2006; Kamienkowski et al., 2011). This

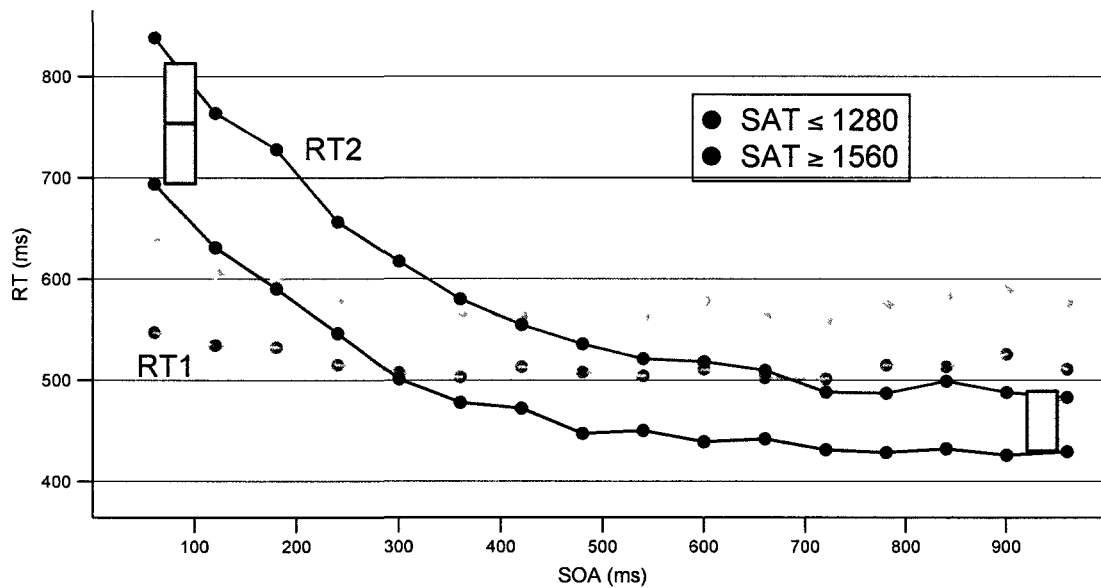


Figure 9: **The locus of the  $g$ -RT correlation is the serial processing stage responsible for the PRP.** The difference in RT2 associated with  $g$  doubles as the SOA diminishes to 60 ms. This indicates that the  $g$ -associated stage cannot be executed concurrently for two stimuli presented close together in time. The height of the gold bar corresponds to the asymptotic difference in RT2 between the moderate- and high- $g$  groups (57 ms). Compare with Figures 7F and 8F.

executive stage also contains low-variability and stochastic components, which gives it a fractal resemblance to the entire multistage computation in which it is embedded. This stage can be prolonged by shortening the SOA in conditions without specific instructions regarding which stimulus should be responded to first. Although our own instructions called for a strict prioritization of the first stimulus, we still found a slight dependence of the RT1 delay on SOA (Figures 8 and 9). It is possible that our use of the same task twice may have exacerbated the scheduling conflict at short SOAs.

Both the moderate- and high- $g$  groups showed a slight increase in RT1 as the SOA increased past 720 ms. This delay at long SOAs bears the signature of prolonging the  $M$  stage: a rigid shift in the entire distribution of RT1 that is not propagated into RT2. We speculate that the SOA affected muscle tension in the left hand (Sanders, 1990, 1998), although it is unclear why our experimental design would produce such an effect. Looking at SOAs between 360 and 720 ms, we found that the RT1 delay was virtually identical in the two  $g$  groups (71 ms). However, as the SOA decreased below 360 ms, the increase in RT1 was steeper in the moderate- $g$  participants ( $\beta_{\text{SOA}} = -.23 \pm .02$ ;  $\beta_{g \times \text{SOA}} = .078 \pm .03$ ). This interaction between  $g$  and SOA resulted in an RT1 difference at the shortest SOA of 87 ms, which is about 30 ms more than the difference in RT1 (and RT2) observed asymptotically. Additionally, whereas the variance of RT1 was roughly constant across the entire range of SOA in the high- $g$  participants ( $\beta_{\text{SOA}} = -2.8 \pm 1.9$  ms), the variance of RT1 increased sharply in the moderate- $g$  participants as the SOA diminished below 360 ms ( $\beta_{\text{SOA}} = -64 \pm 16$  ms;  $\beta_{g \times \text{SOA}} = 57 \pm 23$  ms).

These observations are most simply explained by broadening our motivating hypothesis: the executive and central operations preceding a dual-task response constitute a serially

traversed decision tree, each node branching to one of a few discrete alternatives, and  $g$  is associated with the rate of the diffusion implementing the computation at each node. Asymptotically the scheduling conflict poses little burden to the central system, and thus the RT1 delay in this temporal regime mainly reflects the low-variability contribution to the  $E$  stage. Since  $g$  is not associated with this contribution, the RT1 delay does not interact with  $g$  in the asymptotic regime. As the scheduling conflict worsens with decreasing SOA, however, the decline in the “baseline” rate of the diffusion within  $E$  allows any additional decrement associated with lower  $g$  to show up as noticeable increases in the duration and variability of the time spent deciding which stimulus should be prioritized for central processing.

The calculation of the factor by which the  $g$  difference in RT2 increases with diminishing SOA must adjust for the concomitant  $g$  difference in  $E$ , which is propagated into RT2 in this temporal regime. After this adjustment we found that reducing the SOA into the interference regime increased the difference in RT2 between the moderate- and high- $g$  groups by the factor 1.99 [99% CI = (1.21, 4.02)]. We clearly cannot reject the hypothesis that the factor is precisely equal to two. In contrast, we reject the hypothesis that the factor is equal to unity ( $p < .005$ ). Our data thus support the hypothesis that  $g$  is associated with higher accumulation rates within each of several serially arranged decisional stages.

Our repetition of the same task to constitute our dual task allowed us to decompose total RT into estimated durations of  $P + M$  and  $C$  (Materials and Methods), quantities that are not separately available in most PRP studies. Because these estimates do not rely on any information beyond what we used in calculating the factor by which the  $g$  difference in RT2 increases, they provide additional validation of our hypothesis regarding the nature of the  $g$ -RT association only to the extent that they are consistent with the results of other investi-

gators. We found that the moderate- and high- $g$  groups differed on average from each other in  $P + M$  by less than 1 ms; the mean of the entire sample was 229 ms. Previous studies, using diverse species, tasks, and methodologies, have estimated the visual processing of an easily detectable stimulus to last typically about 200 ms (Dehaene, 1996; Reynolds et al., 1999; Gold & Shadlen, 2000; Lee et al., 2002; Roelfsema et al., 2002; Sigman & Dehaene, 2005; Li et al., 2008). Given the assumption that the average  $M$  is relatively brief, lasting about 30 ms, our partition of RT is in good agreement with those obtained using other approaches. Incidentally we estimated the mean duration of  $C$  to be 268 ms in the moderate- $g$  group and 212 ms in the high- $g$  group.

Another item of evidence regarding the source of the  $g$ -RT association is the correlation across trials of RT1 and RT2. Figure 10 shows that this correlation was initially of similar magnitude ( $\sim .80$ ) in both the moderate- and high- $g$  groups. At an SOA of 240 ms, the correlation began to decline in both groups, but more precipitously in the high- $g$  group. Our simulations showed that such a pattern is compatible with a difference between the  $g$  groups only in the serial diffusion process and not in the other stages (Materials and Methods). When the SOA is 180 ms or less,  $P_2$  almost always finishes while the processing of the first stimulus is still somewhere within  $P_1$ ,  $E$ , or  $C_1$ . Therefore, in this temporal regime, the initiation of  $C_2$  is time-locked to the termination of  $C_1$ , producing the strong RT1-RT2 correlation. Starting at an SOA of 240 ms, however,  $C_1$  is sometimes finished before the termination of  $P_2$ . This is because the stochasticity of the processing within executive and central stages can result in individual hitting times that are shorter than the mean times. When  $C_2$  is free to start immediately, RT1 and RT2 are no longer locked together. The serial portion of RT1 is shorter, less variable, and less right-skewed in the high- $g$  participants, freeing the

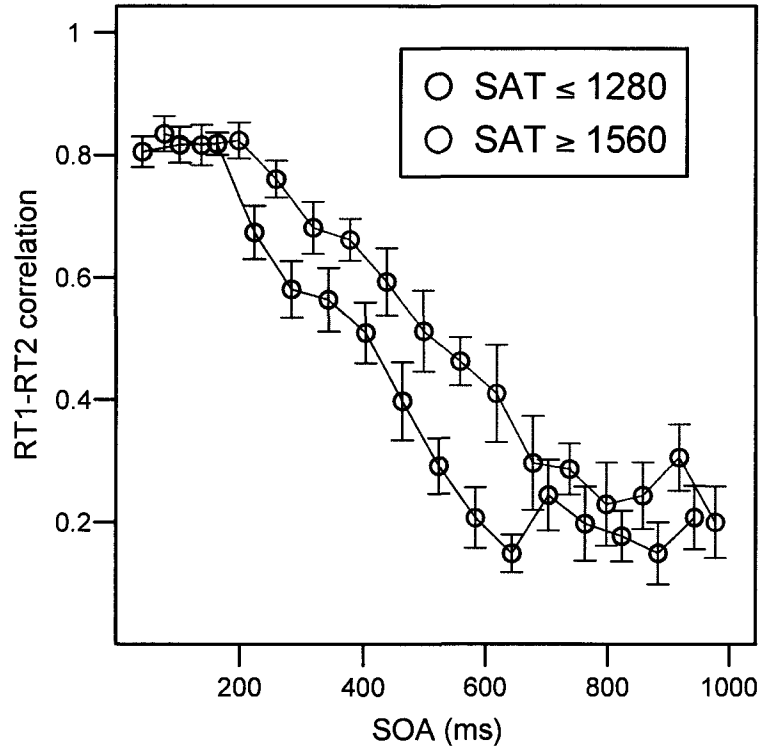


Figure 10: **The zero-lag correlation between RT1 and RT2 as a function of SOA.** For clarity the data points for the moderate- and high- $g$  groups are horizontally displaced by a small quantity. The moderate- $g$  correlation is always slightly to the right of the high- $g$  correlation corresponding to the same SOA. Each point is the average correlation of the participants in the  $g$  group. The bars encompass  $\pm 1$  standard error.

initiation of  $C_2$  on an increasingly greater proportion of their trials, and thus their RT1-RT2 correlation declines more steeply with SOA.

## 2.3 Discussion

We now summarize the contribution of our study and how it meshes with the previous work leading to the unified model depicted in Figure 3.

We can imagine the three phenomena treated in this study—(i) the serial  $C$  stage of the PRP, (ii) the stochastic evidence-accumulation stage of the diffusion model, and (iii) the trait of general intelligence—as corresponding to the three vertices of a triangle. By showing that the stochastic evidence-accumulation stage of the diffusion model is contained *within* the  $C$  stage of the PRP model, Sigman and Dehaene (2005) have connected (i) and (ii). By showing that  $g$  is correlated only the evidence-accumulation stage of the diffusion model and not with the residual time, Schmiedek et al. (2007), Wagenmakers (2009), and Ratcliff et al. (2010) have connected (ii) and (iii). There is a missing side of the triangle, namely the connection between (i) and (iii), but the emerging unified theory demands the existence of this side. Our experiments have provided an array of results upholding this deduction.

We first conducted an additive-factors RT experiment manipulating stimulus-background contrast, the numerical distance of the stimulus from the reference, and response finger. The only statistically significant interaction among  $g$  and the three manipulations was that between  $g$  and numerical distance, supporting the following two inferences: (1) each manipulation affects a distinct stage, and (2)  $g$  is associated only with the stage affected by numerical distance. A diffusion decomposition of the data from this same experiment showed that whereas  $g$  and numerical distance were both associated with diffusion rate, contrast

and response finger were associated with residual time. The additive-factors and diffusion approaches thus converged in singling out the one stage affected by numerical distance as having a privileged relationship with  $g$ . We then performed a classical PRP experiment to determine how a  $g$ -associated difference in RT changes with increasing interference between stimuli. If this difference resides entirely in a single stage *and* that stage is serial (i.e., corresponding stages for different stimuli cannot proceed simultaneously), then the difference must double as the SOA becomes very small. This is because the stage processing the second stimulus must await the completion of the stage processing the first stimulus. If the difference between two individuals in single-task or asymptotic RT is attributable entirely to this stage, then the slower individual is penalized twice in a dual task with a short SOA. We did in fact observe the predicted doubling of the  $g$ -associated RT difference as the SOA diminished, confirming our deduction that  $g$  is associated with the speed of a serial processor. The more precipitous decline of the RT1-RT2 correlation with SOA in the high- $g$  group was also consistent with this deduction

To summarize, we have shown that the  $g$  advantage in RT does indeed reside in the serial stage that has been shown to encompass the stochastic contribution to RT. The mapping of retinal input to an abstract quantity representation, which can be executed with little trial-to-trial variability and in parallel for at least two stimuli, is not associated with  $g$ . The implementation of the selected motor response is also a low-variability, parallel stage with no detectable association with  $g$ . In contrast, the central stage contains a stochastic accumulation of evidence regarding the stimulus-response mapping, and the positive correlation of the accumulation rate with  $g$  is what accounts for the overall  $g$ -RT association. In more complex tasks, there are several distinguishable decisional stages intervening between perception and



action, and it appears that  $g$  is positively correlated with the accumulation rate of each such stage.

We now point out a caveat regarding the relationship of  $g$  with the  $M$  stage. Fine-grained tracking of limb movements in RT tasks requiring participants to reach for the response key have revealed that central manipulations have effects on the trajectory of the reach (Song & Nakayama, 2009). These observations refute the constant-output axiom with respect to  $C$ ; the effect of numerical distance in the number-comparison task continues to “leak” into  $M$  while the latter stage is in progress. Thus, although our additive-factors experiment did not yield any statistically significant interactions among  $g$ , numerical distance, and response finger, a better-powered study may conceivably find these interactions to be weak but reliable. A dynamic interpenetration of the  $C$  and  $M$  stages may explain the correlation between  $g$  and movement time in some previous studies (Jensen, 1987a,b), the complex effects of handicapping the response effector on diffusion parameters (Voss et al., 2004), and the effect of finger on diffusion rate ( $v$ ) in our own data. Perhaps an apt analogy is the interleaving of screen-writing and principal photography despite the conceptual distinction between these stages in the production of a motion picture. Despite this breakdown of a constant-output model, we nevertheless note previous successful attempts to decompose RT into distinct  $C$  and  $M$  stages (Sanders, 1990, 1998; Pashler, 1998; Sigman & Dehaene, 2005). For instance, in a word production task, the selection of a lemma (abstract mental representation of a word) occurs in  $C$  while the selection of phonemes occurs in  $M$  (Ferreira & Pashler, 2002). The positing of discrete  $C$  and  $M$  stages, with  $g$  being associated only with the former, may approximate the time course of processing reasonably well when the task requires a punctate motor response such as a key press. Furthermore, future studies may turn the breakdown of this model out-

side its domain of applicability into an asset. That is, it may be possible to use the precise spatiotemporal trajectories of continuous movements in motorically demanding RT tasks to shed further light on the nature of central processing and  $g$ .

In previous research it was found that the dependence of diffusion rate on numerical distance in the number-comparison task was linear over ranges of numbers similar to those used in our additive-factors experiment (Dehaene, 2007). In our own data, however, there was a shallower increase in diffusion rate with distance for numbers relatively far from the reference (Figure 6B). This result may be partly an artifact of our method for estimating diffusion rates, which suffers from a downward bias that increases with the rate itself (Wagenmakers et al., 2007; van Ravenzwaaij & Oberauer, 2009; Grasman et al., 2009). But as a logical matter we might expect a curvilinear relation between distance and diffusion rate because it is implausible that the rate should continue to increase linearly as the distance becomes very large. The curvilinear relation revealed by our data may be attributable to the unusually high average level of  $g$  in our sample, which must have resulted in a baseline diffusion rate closer to the asymptote. This suggests that individual differences can provide their own domain of applicability for quantitative laws intended to describe species-typical behavior.

An interesting ancillary result is that the estimated duration of  $P + M$  was about 100 ms shorter than the estimated mean duration of  $T$  in Table 10. Since the known bias in the EZ-diffusion methods probably cannot account for this discrepancy, it is likely that  $T$  and  $P + M$  are not coextensive. That is, there are substages within the serial  $C$  stage other than the diffusion process. Although the neural and psychological properties of these substages remain to be elucidated, it is already apparent that these substages are not strongly associated with  $g$ .

The importance of having isolated the  $g$ -RT association to a particular stage of course

depends on the causal nature of the association. It is conceivable that the  $g$ - $C$  association arises entirely as a result of trivial confounding or reverse causation. For example, diffusion rates do increase with practice (Dutilh et al., 2009; Kamienkowski et al., 2011), and it may be that higher- $g$  individuals perform central processing in the number-comparison task more rapidly simply because they have been exposed more to numbers. There is already reason to think, however, that trivial confounding and reverse causation cannot fully explain the  $g$ - $C$  link. The negative  $g$ -RT correlation has been found to hold within families (Jensen et al., 1989), which rules out many conceivable sources of confounding (Jensen & Sinha, 1993; Beauchamp et al., 2011). Also, higher- $g$  individuals tend to respond more rapidly even in very simple tasks such as detecting the onset of a single light (Jensen, 1987a), where it seems unlikely that  $g$  could be associated with greater practice or familiarity. An urgent priority for future research is to verify the presumed causal effect of central processing on  $g$ , perhaps drawing upon the powerful observational designs for causal inference that have been developed in recent years (Spirtes et al., 2001; Gillespie & Martin, 2005; Pearl, 2009).

Because the unified PRP-diffusion model has interfaces with both algorithmic and neural levels of analysis (Marr, 1982), establishing the causal effect of central processing (diffusion rate) on  $g$  would connect the study of individual differences to multiple lines of reductionistic investigation. One such line would attempt to incorporate a hierarchy of serial, stochastic decisional stages in models of more complex and naturalistic thought processes (Carpenter et al., 1990; Hofstadter & the Fluid Analogies Research Group, 1995; Johnson-Laird, 2006; Sackur & Dehaene, 2009). Indeed, one avowed aim of researchers developing the unified model is to elucidate how algorithmic processing can be enabled in neural machinery (Dehaene, 2008).

A research program beginning with the unified model in Figure 3 may aim not only upward at algorithmic decompositions of thought, but also downward at low-level properties of the brain. Combining electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), recent studies have spatially circumscribed the temporal signature of the serial *C* stage to a suite of frontoparietal regions (Dell'Acqua et al., 2005; Marois & Ivanoff, 2005; Sigman & Dehaene, 2008; Hesselmann et al., 2011). Further improvements in resolution are necessary to fit the data from imaging studies of the PRP to detailed models of neuronal activity. Encouragingly the *C* regions, as now delimited, coincide in broad outline with those singled out by imaging studies of *g* (Jung & Haier, 2007).

The stochastic nature of the evidence accumulation within the *C* stage constitutes another distinctive signature. Data from both monkeys and humans indicate that the accumulation is implemented by a transient recurrent network of sensory, parietal, prefrontal, and motor regions recruited specifically for the given task (Gold & Shadlen, 2007; Heekeren et al., 2008), and future work may bring about a convergence of *C* and diffusion correlates in a core frontoparietal region. Across various tasks this core may bind together the appropriate peripheral processors to constitute the “blackboard” or “global workspace” posited in theories of high-level cognition. The mechanism by which this core network might implement a diffusion of evidentiary strength between decision boundaries is not yet known. One appealing model has treated the outputs of two reciprocally inhibiting pools of neurons as a dynamical system with a saddle point separating the basins of attraction corresponding to the response alternatives (Wong & Wang, 2006). In this model the timescale of the diffusion follows from well-known properties of the NMDA glutamate receptor. It would be worthwhile to investigate whether the positive correlation between *g* and diffusion rate can

be reproduced by varying plausible graph-theoretic analogs of  $g$ 's known neural correlates, including network order and connectivity, in a biophysically realistic non-reduced form of the model representing individual nodes within each pool.

## **2.4 Materials and Methods**

### **2.4.1 Participants**

Recruitment at various universities in the Boston area began in January 2011 and continued until the end of May. One high- $g$  participant who completed the additive-factors experiment was removed from the dataset for being unable to see the gray stimuli from a distance greater than a few cm. This participant showed a main effect of the contrast manipulation exceeding 120 ms, four times the effect estimated from the other participants, and statistically significant interactions of contrast with other manipulations. Another high- $g$  participant who completed the additive-factors experiment was removed for excessive limb and body movement (swiveling in her chair, swaying from side to side, crossing her legs and shaking her foot). This participant showed an uninterpretable pattern of significant crossover interactions. These participants were flagged for removal before their data were analyzed.

Overall a total of 51 individuals participated and were not removed. Twenty-one individuals (12 high  $g$ , 9 moderate  $g$ ) participated in the additive-factors experiment; twenty-eight individuals (14 high  $g$ , 14 moderate  $g$ ) participated in the dual-task experiment; two individuals who were not screened for  $g$  participated in the dual-task experiment testing the parallelism of the stage affected by stimulus-background contrast. The mean age of the high- $g$  participants was 21.1 years with a standard deviation of 2.76; the mean age of the moderate-

*g* participants was 20.8 years with a standard deviation of 1.64. Participants were all fully fluent English speakers and were paid a base of forty dollars for their time.

#### **2.4.2 Procedure**

Participants performed the number-comparison task and completed an IQ test in a session lasting between 2 and 3.5 hours. Up to 3 participants were run under the supervision of an experimenter in a single session. Each participant was seated at a cubicle obstructing all other participants from view. All participants in a session were members of only one *g* group. Assignment to cubicles was randomized.

The moderate-*g* participants were told that they were participating in “Version A,” while the high-*g* participants were told that they were participating in “Version B.” After completing all tasks, each participant had to predict the outcome of a die roll. Within each version (A or B) of each experiment (additive factors or dual task), all participants who correctly predicted the roll were entered in a pool and randomly paired. Within each such pairing, the participant who performed better on the sum of mean RT on correct trials in the number-comparison task and number correct on the IQ tests was mailed a check for an additional 40 dollars. Each term in this sum was standardized within each version-experiment combination and appropriately signed. Sacrificing accuracy for speed was discouraged by penalizing accuracy below 90 percent with a multiplication of mean RT by  $(1+.20x)$ , where  $x$  is the participant’s number of rounded percentage points below 90-percent accuracy. This incentivization of good performance was explained, with a numerical example, to participants at the beginning of the session.

The additive-factors experiment balanced and randomized the two levels of stimulus-

background contrast, the 48 possible numbers, and the two levels of response finger; each combination occurred 8 times. These 1536 actual trials were broken up into 32 blocks of 48 trials each. Contrast and numerical distance varied within blocks, whereas finger varied across blocks. Participants responded either by pressing the R and U keys with their index fingers or by pressing the W and O keys with their ring fingers. The interval between trials was 750 ms. Participants familiarized themselves with the task over a total of 154 practice trials before beginning the actual experiment. Because the precise effect of contrast depends on viewing angle, participants were instructed to adjust seat height and monitor angle during the practice trials in order to maximize the visibility of the gray stimuli.

The dual-task experiment used each of the 64 possible combinations of numbers 16 times, once for each SOA (60 to 960 ms inclusive, increments of 60 ms). After randomization these 1024 real trials were broken up into 32 blocks of 32 trials each. Participants responded by pressing the Q and W keys with the middle and index fingers of their left hands and the O and P keys with the index and middle fingers of their right hands. The interval between trials was 1000 ms. Participants familiarized themselves with the task over a total of 106 practice trials before beginning the actual experiment. One practice block consisted of 32 trials, all with an SOA of 60 ms, in order to accustom participants to performing the task quickly and accurately without “grouping” responses (withholding the first response and then emitting both responses in a rapid burst).

The version of the dual-task experiment varying contrast followed the same design as just described, except that there were only 4 SOAs (50 to 950 ms inclusive, increments of 300 ms). All 256 possible combinations of numbers and contrast level were used 6 times to constitute the real trials.

In both experiments participants were told to keep their fingers in contact with the keys at all times during the blocks and to minimize all limb and body movement. After every fourth block, participants were given an indefinitely long break, during which they were allowed to use the restroom, have a cup of water, and so on. The next block was initiated when all participants indicated that they were ready. All other inter-block intervals were 20 s in duration.

After completing the number-comparison task, participants were administered a short form of Raven's Advanced Progressive Matrices and the vocabulary subtest of the Multidimensional Aptitude Battery II. No time limit was imposed.

### **2.4.3 Stimuli**

Both the additive-factors and dual-task experiments were implemented in PsyScope X Build 57 and run on iMac desktops (Mac OSX Version 10.5.8, 2.66GHz Intel Core 2 Duo, 4GB 1067 Mhz DDR3). The diagonal length of the monitor was 50.8 cm.

The stimuli were displayed in the Monaco font with a point size of 36. The distance from the monitor to the eyes of a given participant was roughly 70 cm. In the additive-factors experiment, we employed the PsyScope default for the black stimuli and the setting (−5000, −5000, −5000) for the gray.

### **2.4.4 Data Analysis**

All trials resulting in RT (or RT1) less than 250 ms were discarded. All additive-factors trials resulting in RT more than 4 standard deviations from a participant's mean in a given cell (contrast, numerical distance in eight bins, response finger) were discarded iteratively.



All dual-task trials resulting in RT more than 4 standard deviations from a participant's mean in a given cell (task order, SOA, dichotomized numerical distance) were discarded iteratively. These criteria eliminated about 1 percent of all trials in both the moderate- and high- $g$  groups.

In none of the analyses below did using the continuous measurements of IQ produce substantively different results from dichotomized  $g$  level. The failure to obtain greater signal with the continuous measurements may be due to selection bias producing a negative correlation between  $C$  speed (diffusion rate) and the other causes of  $g$  within each group (Pearl, 2009).

Linear mixed models were fit to the additive-factors data using R's lmer package. All confidence intervals and  $p$ -values for other analyses were  $BC_a$  intervals calculated using the boot and simpleboot packages. In cases where the  $BC_a$  interval agreed well with the normal-theory interval and the  $p$ -value was less than .001, normal theory was used to calculate the  $p$ -value.

Let  $X$  denote the random hitting time of a Wiener diffusion process. The defective cumulative probability of error hitting times is

$$P(\text{error}, X \leq x) = P(\text{error}) - \frac{\pi s^2}{a^2} \exp\left(-\frac{zv}{s^2}\right) \sum_{k=1}^{\infty} \frac{2k \sin\left(\frac{\pi kz}{a}\right) \exp\left[-\frac{1}{2}\left(\frac{v^2}{s^2} + \frac{\pi^2 k^2 s^2}{a^2}\right)(x - T)\right]}{\left(\frac{v^2}{s^2} + \frac{\pi^2 k^2 s^2}{a^2}\right)}, \quad (1)$$

where

$$P(\text{error}) = \frac{\exp\left(-\frac{2av}{s^2}\right) - \exp\left(-\frac{2zv}{s^2}\right)}{\exp\left(-\frac{2av}{s^2}\right) - 1} \quad (2)$$

is the probability of absorption at the wrong boundary. To obtain the equation that gives the

defective probability of a correct response before  $x$ ,  $z$  and  $v$  should be replaced by  $-z$  and  $-v$  respectively in Equation 1. In these expressions  $T$ ,  $a$ , and  $v$  are defined as in Table 7.  $s^2$  is a scaling constant that is conventionally set to 0.1.  $z$  is the starting point of the process. Given the symmetry of our number-comparison task, we may set  $z = a/2$ . A derivation of these expressions is given by Feller (1968), who takes the limit of a discrete random walk with probability  $p$  ( $q$ ) of heading upward (downward) at each epoch as  $p - q$  becomes small, the size of each step becomes small, and the number of epochs per time unit becomes large. This approach is useful because it immediately shows why the variance of the process increases with the mean. Since the variance of a dichotomous random variable is maximized at  $p = 1/2$ , a smaller value of  $p - q$  means greater variability as well as a shorter average distance traveled during a fixed number of epochs.

Given the limited number of trials in any given cell and the high accuracy of our participants, we were not able to fit Equations 1 and 2 to individual data (Vandekerckhove & Tuerlinckx, 2008). Attempting to fit the equations to averaged data did not produce satisfactory results, possibly because the nonlinearity of the diffusion model precludes the parameters estimated from averaged data necessarily converging on the average values of the parameters in the sample. The thresholds used to recruit our samples may have exacerbated this problem. Remarkably there exist closed-form expressions for the means and variances of the distribution defined by Equations 1 and 2 that permit  $T$ ,  $a$ , and  $v$  to be estimated by the method of moments (Wagenmakers et al., 2007; van Ravenzwaaij & Oberauer, 2009; Grasman et al., 2009). This approach relies on the assumption that  $T$ ,  $v$ , and  $z$  do not vary across trials within a cell for a given participant. Although variability in these parameters is required to fit the distribution of error RT (Ratcliff & Smith, 2004; Ratcliff & McKoon,

2008), it has been found that an incorrect assumption of no variability still permits the recovery of experimental effects and individual differences. One reason for the robustness of the reduced moment-based method is that the variance of the diffusion process is empirically an order of magnitude larger than the variance of  $T$ .

Fitting the non-reduced diffusion model to the entire distribution of RT would impose strong constraints on not only the behavior of the RT2 central tendency as a function of SOA but also on several other RT2 quantiles. This would arguably provide an even more demanding test of the model in Figure 3. Such fitting was done successfully in a previous study (Sigman & Dehaene, 2005), but this analysis did not account for errors. A complete diffusion analysis of a PRP task, although requiring many more trials than we administered in our own experiment, would certainly be worthwhile.

A complete version of the model in Figure 3, incorporating the executive task-scheduling stage, can be written as

$$\text{RT2}(\text{SOA}) = \begin{cases} P_1 + E(\text{SOA}) + C_1 + C_2 + M_2 - \text{SOA} & \text{if } \text{SOA} + P_2 \leq P_1 + E + C_1, \\ P_2 + C_2 + M_2 & \text{if } \text{SOA} + P_2 > P_1 + E + C_1. \end{cases} \quad (3)$$

It is evident from this equation that a difference associated with  $C$  must double as the SOA becomes small. Simulations can be performed by assigning a probability distribution to each term ( $E$  and  $C$  each being the convolution of a low-variability substage and a diffusion process approximated by a discrete random walk). Direct estimates of  $P + M$  and  $C$  are made possible by assuming that corresponding stages of tasks 1 and 2 have the same mean, which is plausible if tasks 1 and 2 are identical. Then we may simply take  $\mathbb{E}[2 \times$

$RT1(SOA) - RT2(SOA) - E(SOA) - SOA]$  for short SOAs, which by Equation 3 is equal to  $\mathbb{E}(P + M)$ .  $\mathbb{E}[E(SOA)]$  may be estimated by  $\mathbb{E}[RT1(SOA) - RT2(\infty)]$ . We used this method, averaging the results from the 60- and 120-ms SOAs, to estimate  $\mathbb{E}(P + M)$  separately for the moderate- and high- $g$  groups.

## **Acknowledgments**

We thank Steven Pinker for helpful discussions.

### **3 Correlation and Causation in the Study of Personality**

James J. Lee<sup>1</sup>

**1 Department of Psychology, Harvard University, Cambridge, MA, USA**

#### **Abstract**

The aim of personality psychology is to explain the causes and consequences of variation in behavioral traits. Because of the observational nature of the pertinent data, however, this endeavor has attracted much controversy. In recent years the computer scientist Judea Pearl has used a graphical approach to extend the innovations in causal inference developed by the population geneticists Ronald Fisher and Sewall Wright. Besides shedding much light on the philosophical notion of causality itself, this graphical theory now contains many powerful concepts of relevance to the controversies just mentioned. In this article some of these concepts are applied to areas of personality research where questions of causation arise, including the analysis of observational data and the genetic sources of individual differences.

Keywords: personality; causality; directed acyclic graph; structural equation modeling; behavioral genetics

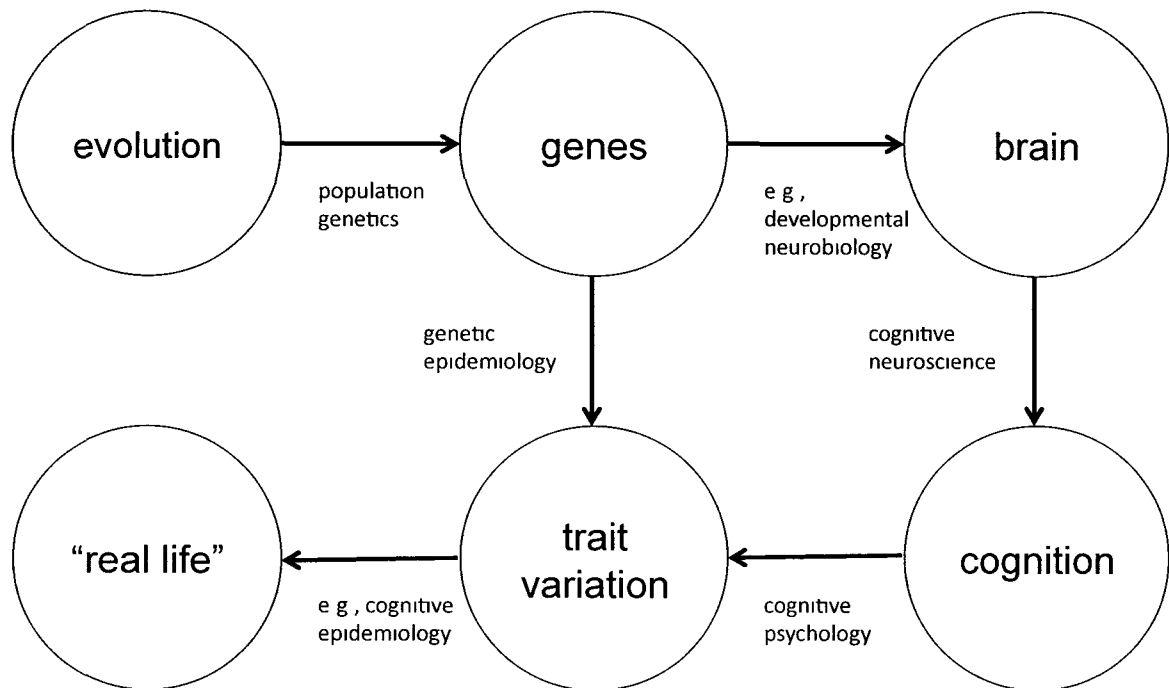
## Acknowledgements

Please send correspondence to jameslee@wjh.harvard.edu. I am particularly grateful to Allan Drummond, Tom Bouchard, and Judea Pearl for their encouragement and generosity.

Until recently the notion of *causality* remained a merely qualitative concept, subject to the hostility of mathematicians such as Karl Pearson and Bertrand Russell. Under this limitation it was impossible to find a formal notation to express even such simple notions as *rain causes mud and not vice versa*; in merely probabilistic terms, the most we could say was that *rain and mud are correlated*. Building on the foundations laid by the population geneticists Ronald Fisher and Sewall Wright, the computer scientist Judea Pearl and his colleagues have filled this lacuna in scientific discourse with a simple yet powerful formalization of causality that draws on the branch of mathematics known as *graph theory*. Pearl's axiomatization of causality stands to offer a particularly great benefit to the study of personality, where for various reasons (not all strictly scientific) the difficulties of pursuing causal claims without a respectable causal vocabulary have been particularly keen. Indeed, despite the difficulty in interpreting the proposed chain in Figure 11 as anything but a *causal* chain, the writings of Pearson, Russell, and other Edwardian scientists have sometimes persuaded personality theorists to deny that causality is what they are trying to demonstrate (Burt, 1940; Lubinski & Dawis, 1995).

The thesis of this article is that Pearl's graphical theory provides a foundation for the properly causal ontological commitments of practicing personality researchers. In Part 1

Figure 11: Causal chain hypothesized by some personality psychologists. This chain happens to be a directed acyclic graph, although it does not represent any formal causal model. Only a subset of the possible nodes and edges is depicted.



I demonstrate the power of the theory in several examples representative of the problems arising in personality research. These examples shed light on the following general issues:

- *Correlation and causation.* We often hear the mantra *correlation does not equal causation*. But what then *is* the relation between correlation and causation? According to the graphical theory, every non-coincidental correlation arises from some causal mechanism, perhaps involving variables other than the pair under consideration. The graphical theory thus provides a fundamental taxonomy for classifying correlations according to the causal structures that have generated them.
- *Covariate selection.* We are often told to control for potential confounders in an observational study by including them as regression covariates. But what exactly is a confounder? That is, how do we decide which variables to control for? Is there ever a reason *not* to control for some measured variable?
- *Randomization.* A frequent justification of randomized assignment to different levels of the putative causal variable invokes the tendency of randomization to create groups that are similar in background characteristics. While this argument is valid, it may become less obviously so after the discussion of Topic 2. If the graphical theory of causality is truly comprehensive, it should be able to supply its own justification for randomizing participants to different levels of the putative causal variable whenever this is feasible.

In Part 2 I take a necessary digression to discuss the nature of psychometric common factors—the very objects of study in personality research. A frequent objection to the scientific status of personality research is that *g*, the Big Five (or Six) personality traits, and



other factor-analytic “constructs” are arbitrary mathematical fictions (Gould, 1981; Glymour, 1997). This objection is often part of a longer argument: since factor analysis is hopelessly inadequate as a tool of causal discovery, any scheme that supposes psychometric common factors to be meaningful causes or consequences of other variables must be similarly unsound. Part 2 gives a minimal argument countering this kind of nihilism. Although I also deny that a psychometric factor stands in a causal relation to its indicators, I do allow a factor to play the role of cause or effect in a larger network.

The discussion in Part 1 will compel the conclusion that *structural equation modeling* (SEM) is inevitably employed whenever investigators advance a causal claim on the basis of observational data. Accordingly, in Part 3 I reanalyze a dataset bearing on the relation between intelligence and social liberalism in order to demonstrate how Pearl’s graphical approach can sharpen the explicit use of SEM in personality research.<sup>1</sup> In particular, the graphical approach provides a means of identifying the testable implications (if any) of causal hypotheses for observational data. We can then subject these implications to severe empirical tests, the survival of which can confer great credibility to (a class of) causal models even in the absence of randomization. Here the contrast with mainstream SEM could not be greater. Some psychometricians go so far as to say that the purpose of SEM is not to shed light on causation but rather to express conditional probability distributions in a different form (Muthén, 1987; Holland, 1995; Schumaker & Lomax, 2004). In the authoritative reference on psychometrics in the *Handbook of Statistics* series, the entry on SEM contains excellent coverage of statistical issues but no discussion of causality at all (Yuan & Bentler,

---

<sup>1</sup>Trent Kyono has written a beta version of the program Commentator, which automates many of the analyses demonstrated in Parts 1 and 3. Email him at [tmkyono@gmail.com](mailto:tmkyono@gmail.com).

2007). An important aim of this paper is to restore Wright's (1968) view of SEM as an exercise in applying the logic of cause and effect.

In Part 4 I take up the intersection of graphical methods and an emerging research area of vital importance to the entire structure depicted in Figure 11: the search for DNA polymorphisms with causal effects on personality. The cost of sequencing the entire genome of a research participant will eventually be negligible, and at that point gene-trait association research may succeed brain imaging as the "land grab" of behavioral science. Such research on diseases and anthropometric traits has already yielded spectacular dividends; one recent study of over 180,000 participants uncovered 180 genomic regions containing a variant affecting height (Lango Allen et al., 2010).

Since the nature-nurture issue has been a flash point in the controversies that have dogged personality research, this article's commitment to the utility of genetic research may seem inauspicious. Here I give two related reasons for concluding my article in this way. First, population genetics now contains many theoretical results developed without the benefit of a general framework for causal reasoning, and the new explanations of these results inspire confidence in the unity and generality of the graphical approach. Second, many of the examples preceding Part 4 will show that causal inferences can depend on assumptions that are untestable given the data at hand. For instance, the discussion in Part 3 invokes temporal ordering to rule out alternative causal models, but this assumption is admittedly fraught. A latent developmental process may predetermine  $Y$  well before  $X$ , even if  $X$  is the first trait to be manifest in the lifespan. The upshot is that the soundness of any causal conclusion depends on both conforming data and the correctness of the requisite assumptions. Our substantial prior knowledge of genetics allows us to justify many powerful assumptions,

which leads to correspondingly powerful results. Gene-trait association research thus provides many enlightening applications of graphical reasoning:

- *The meaning of heritability.* As documented by Sesardic (2005), immense confusion persists over the meaning of heritability and the nature of the evidence supporting heritability estimates for particular traits. Elucidating the notion of heritability from first principles, we will see that causality is built into the very concept. I provide new and simple proofs of Fisher's expressions for heritability and also the Fundamental Theorem of Natural Selection (using, ironically, the path-tracing rules devised by Fisher's bitter rival Sewall Wright). Although these proofs employ less general hypotheses than strictly necessary, they clearly reveal the *causal* content of these expressions.
- *Linkage disequilibrium.* Two genetic loci are in *linkage disequilibrium* (LD) if they are correlated—that is, if knowing a person's genotype at one locus gives some information regarding the genotype at the other. This population-genetic terminology is rather unfortunate in that it applies even to loci not physically linked on the same chromosome, but here we abide by convention. Population geneticists have shown that various ancestral processes, including assortative mating and natural selection, will lead to LD (Fisher, 1918; Bulmer, 1971; Bürger, 2000). The mathematical soundness of these results are not in doubt, but intuitive understanding may be elusive without a graphical interpretation.

For instance, we readily see that spouses tend to resemble each other in some ways. Remarkably, it seems that many of us have absorbed this conspicuous fact of social life without realizing that our intuitive explanation for it (people preferring mates

with certain qualities) does not correspond to anything in the canonical taxonomy of reasons for why any two variables might be correlated. One mate's trait value does not affect the other mate's value, and the two trait values are not confounded in the usual sense. The classification of a so-called marital correlation must invoke a crucial addition that Pearl has made to the correlational taxonomy.

- *Gene-trait association and causation.* Hundreds of genetic loci have now been shown to be associated with one or more traits in genome-wide association studies (Manolio et al., 2009), and most knowledgeable geneticists are confident that the bulk of these associations reflect linkage with authentic causal variants. What justifies this confidence? At first glance the relevant methods (regression, within-family designs, principal components analysis) do not seem so different from those used in other fields, where there have been few uncontroversial causal inferences made on the basis of purely observational data. Using concepts developed in the discussion of the previous topics, we will see how certain special features of gene-trait association studies enable the leap from association to causation.

Some readers may be skeptical that a novel and unified framework can usefully touch on all of the topics just mentioned. A single framework can be so all-embracing, however, precisely because causality is a such a deep and essential concept. It is the burden of this article to demonstrate this depth and essentiality.

## 3.1 A Unifying Theory of Causality

Over the last two decades, Pearl and his collaborators have developed a nonparametric form of SEM that includes the usual idealization of linear causal relations and normally distributed disturbances as a special case. Sprites, Glymour, and Scheines (2001) and their collaborators have also made seminal contributions, although their focus is much more on the automatic generation of causal models. As already emphasized, the importance of this work goes far beyond the extension of traditional SEM.

### 3.1.1 The Interpretation of a Causal DAG

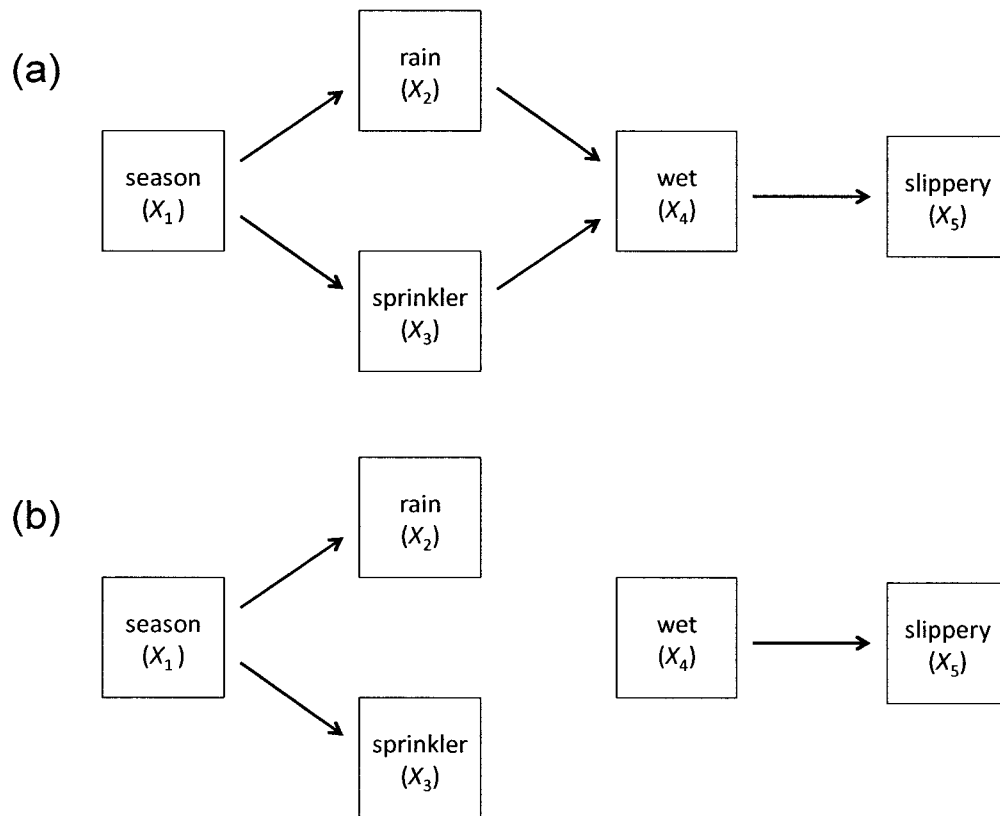
We now review the features of the graphical theory needed in our account of personality and causation. Much of our review focuses on Figure 12, which depicts an example given by Pearl (2009, p. 15). The graph represents the causal relations among five variables: the season of the year ( $X_1$ ), whether it rained last night ( $X_2$ ), whether the sprinkler was on last night ( $X_3$ ), the wetness of the pavement ( $X_4$ ), and the slipperiness of the pavement ( $X_5$ ).

**Definitions and Elementary Properties** The object in Figure 12 is a *directed acyclic graph* (DAG), consisting of discrete *nodes* or *vertices*, some pairs of which are connected by *directed edges*.<sup>2</sup> A *path* is a consecutive sequence of edges with distinct nodes; the edges in a path need not all face the same direction. Readers already familiar with SEM will recognize Figure 12 as a *path diagram* with no representation of the disturbances. In Pearl's theory, however, much greater use is made of such a diagram's formal properties.

---

<sup>2</sup>The graphical theory of causality can accommodate *cycles* representing mutual causation ( $X \rightarrow Y \rightarrow \dots \rightarrow X \rightarrow Y \rightarrow \dots$ ). This paper will not address cyclic models; the reader is directed to Dickens and Flynn (2001) for an example.

Figure 12: A DAG representing the causal relations among five variables (a) before the manipulation of  $X_4$ , and (b) after the manipulation of  $X_4$ .



If there is a directed edge from  $X_i$  to  $X_j$ , then  $X_i$  is a *parent* of  $X_j$ . We extend the analogy to kinship in a straightforward way to define *children*, *ancestors*, and *descendants*. Informally, an ancestor is a cause and a descendant is an effect. In our account we will not encounter a deep analysis of these notions in a single place, but rather allow their meanings to emerge over the course of the discussion.

Now consider the reasons for why we might observe an association between two variables. Two reasons are well-known: (1)  $X$  is a cause of  $Y$  or vice versa, or (2) a third variable is a common cause of both  $X$  and  $Y$  (Fisher, 1970). If either  $X$  or  $Y$  is a cause of the other, they are connected by a *directed path*; each arrow in the path points in the same direction. If there are any intermediate nodes between ancestor and descendant along a directed path, they are called *mediators*. In Figure 12 both  $X_4 \rightarrow X_5$  and  $X_1 \rightarrow X_2 \rightarrow X_4$  are examples of directed paths. If a confounding common cause contributes to the association between  $X$  and  $Y$ , there is a path between them that first travels against the arrows to the *confounder* and then travels with the arrows to terminate at the other node. In Figure 12 the subgraph  $X_3 \leftarrow X_1 \rightarrow X_2$  supplies an example of a confounding path: rain and the sprinkler do not affect each other, but they are associated because the season affects both.

Outcomes under manipulation lie close to the heart of what directed paths semantically represent. Suppose that we wrest control of the mechanisms determining  $X_4$  away from nature and set the level of this variable each morning ourselves. We will then find that  $X_5$  continues to depend on  $X_4$  but that  $X_4$  no longer depends on  $X_2$  or  $X_3$ . That is, if we protect the pavement with tarp whenever we are not spraying it with a garden hose, we will find that hosing the pavement is correlated with neither the rain nor the sprinkler. The graphical representation of “overriding nature” in this way is the deletion of all directed edges converging

on  $X_4$  (Figure 12b). The intuition here should be that  $X_4$  is “set free” or “disconnected” from its parents (and other ancestors) once we intervene to determine its value. We must then attribute any persisting associations with other nodes in the graph to these nodes being descendants of  $X_4$ ; in other words,  $X_4$  is either a parent (*direct cause*) or more remote ancestor (*indirect cause*) of any variable with which it remains associated.

Note that whether a variable is a direct or indirect cause of another is always relative to the epistemological situation. In Figure 12 the omission of either  $X_2$  and  $X_3$  would force us to insert a directed edge from  $X_1$  to  $X_4$ . That is, if we were unaware of any mediating mechanism, we would regard the season as directly affecting the wetness of the pavement.

**Experimental and Statistical Control** We have just seen that experimental control amounts to setting a node equal to a constant. Can statistical control be regarded in the same way? If the statistical control takes the form of conditioning on a node along a confounding path, then the distinct effects represented by the terminal nodes are indeed no longer associated. For instance, suppose that the sprinkler has been automated such that it turns on more frequently in drier seasons. In a short time interval during which the sprinkler follows a fixed schedule, when it rains will no longer be associated with when the sprinkler turns on. Thus, if the only non-directed paths between  $X$  and  $Y$  are confounding paths, we simply condition on a set of variables that contains at least one node on each confounding path between  $X$  and  $Y$ . If any association remains between  $X$  and  $Y$ , there must be at least one directed path from  $X$  to  $Y$  representing a causal effect.

Perhaps surprisingly, there are also variables on which we should *not* condition if we want to obtain an unbiased estimate of a causal effect. Earlier we named causation and con-



founding as two reasons for an association between variables. But there is a third reason that seems hardly known at all:  $X$  and  $Y$  may be associated because both are causes of a third variable,  $Z$ , on which we have conditioned. Figure 12 shows how this might occur. Although rain and the sprinkler are independent if we condition on the season, they will become associated once again if we also condition on the wetness of the pavement. That is, if we only observe the pavement on mornings when it is wet, the two causes become *negatively* correlated; knowing that it did not rain *and* that the pavement is wet implies that the sprinkler was indeed activated.

In this situation the variable  $Z$  is a *collider*. We can think of conditioning on a collider as *unblocking* a path that was previously closed to causal flow. Thus, to obtain a clean estimate of a causal effect of one variable on another, the set of covariates (statistically controlled variables) must include a node on each open non-directed path between the two variables, *including* any such paths opened by conditioning on a collider or its descendants. Only then will the remaining open paths between the variables consist solely of causal effects. If we have not conditioned on any colliders, however, we can ignore the paths including them.

The discussion above is encapsulated in Pearl's critical concept of *d-separation*:

A path  $p$  is *d-separated* (or *blocked*) by a set of nodes  $\mathbb{S}$  if and only if

1.  $p$  is a directed path, confounding path, or unblocked colliding path with at least one intermediate non-collider (with respect to  $p$ ) contained in  $\mathbb{S}$ , or
2.  $p$  is a colliding path such that no collider on the path or any of its descendant is contained in  $\mathbb{S}$ .

A path that is not *d-separated* is said to *d-connect* the extreme nodes  $X$  and  $Y$ .

*d-separation* is also defined for pairs of variables. A set  $\mathbb{S}$  is said to *d-separate*  $X$  from  $Y$  if

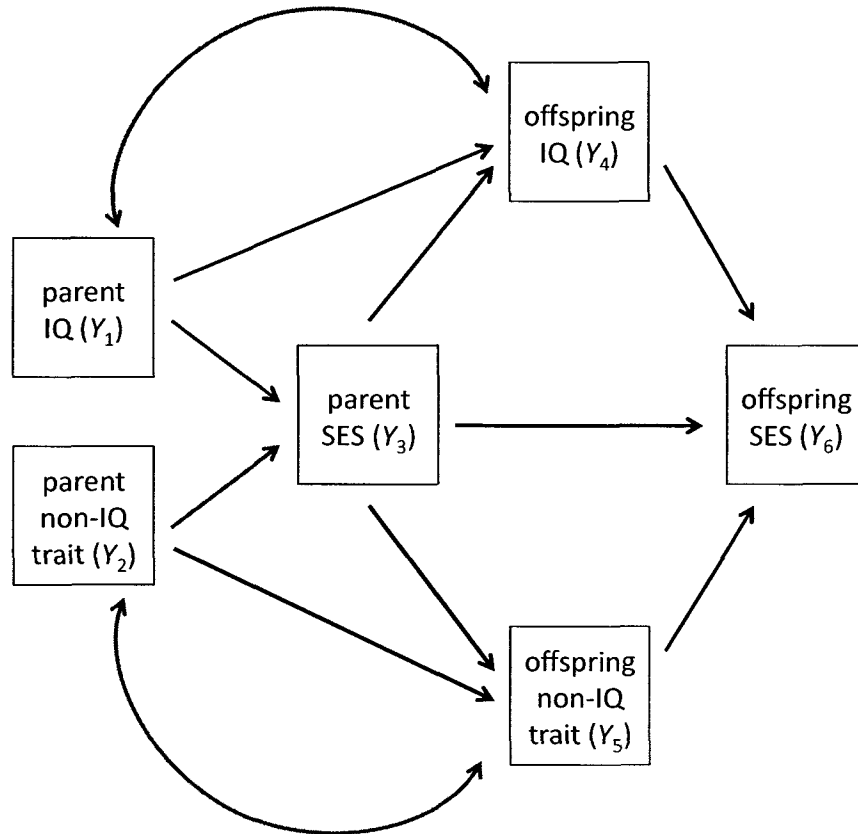
and only if  $S$  blocks every path from  $X$  to  $Y$ . Thus, except in very unusual circumstances, two variables that are  $d$ -connected must be correlated; conversely, any two  $d$ -separated variables must be independent. It will turn out that the broad concept of  $d$ -separation ( $d$ -connection) provides a unifying thread for the topics enumerated in the Introduction.

Colliders demonstrate that conditioning on a variable is not always equivalent to setting it equal to a constant. If we *experimentally* control the wetness of the pavement, sealing off this variable from its natural determinants (including rain and the sprinkler), we are deleting the edges converging on this variable (Figure 12b). This mutilation is unproblematic because the removal of edges can never add a  $d$ -connecting path. *Statistically* controlling the variable, however, merely amounts to examining a subpopulation where all members happen to share the same value. Different members of this subpopulation will have that value for different reasons, which alters the covariation among the variable's causes.

The conceptual distinction between experimental and statistical control motivates Pearl's notational distinction between them. Pearl points out that when statisticians write  $P(Y | X = x)$  to signify the (conditional) probability distribution of  $Y$  given that the variable  $X$  assumes the value  $x$ , they really mean the probability distribution of  $Y$  given that we *see*  $X$  equaling  $x$ . But what scientists want to know is the probability distribution of  $Y$  given that we *do* the action of setting  $X$  equal to  $x$ . The existence of both confounders and colliders shows that  $P(Y | x) = P(Y | \textit{see}(x)) \neq P(Y | \textit{do}(x))$ .

We will now go through two examples showing that heedless conditioning might in fact produce misleading results. Consider the causal model of status attainment, possibly somewhat realistic, in Figure 13. We now incorporate the use of a *bidirectional arc* to represent a residual dependence between two variables attributable to unmeasured common causes. It

Figure 13: A DAG representing a causal model of personality and status attainment.



is not a misnomer to call Figure 13 a directed acyclic graph because a bidirectional arc  $X \leftrightarrow Y$  is simply a shorthand for  $X \leftarrow C \rightarrow Y$ , where  $C$  denotes the unobserved confounders. For simplicity we assume that each variable is well defined and measured without error; in Part 2 I will briefly comment on what these assumptions entail.

The current consensus is that Figure 13 must include the directed edge  $Y_4 \rightarrow Y_6$  (Murray, 2002; Nisbett, 2009). What remains under debate is the relative impact of IQ when

compared to other determinants of SES, including non-cognitive personality traits such as conscientiousness and agreeableness (Roberts et al., 2007). If the SES of the parents is a confounder, the zero-order IQ-SES correlation in their offspring may overestimate the causal effect of IQ. This possibility often motivates including parental SES as a covariate in regression models intended to disentangle the contributions to important life outcomes. Simply including parental SES as a covariate, however, will probably *overcorrect* the estimate of offspring IQ's causal effect. Let  $C_{i,j}$  denote the unmeasured confounders responsible for the bidirectional arc between nodes  $i$  and  $j$ . Conditioning on parental SES  $d$ -separates the confounding paths

$$Y_4 \leftarrow Y_3 \rightarrow Y_6, \quad (4a)$$

$$Y_4 \leftarrow Y_3 \rightarrow Y_5 \rightarrow Y_6, \quad (4b)$$

$$Y_4 \leftarrow Y_3 \leftarrow Y_2 \rightarrow Y_5 \rightarrow Y_6, \quad (4c)$$

$$Y_4 \leftarrow Y_3 \leftarrow Y_2 \leftarrow C_{2,5} \rightarrow Y_5 \rightarrow Y_6, \quad (4d)$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 \rightarrow Y_3 \rightarrow Y_6, \quad (4e)$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 \rightarrow Y_3 \rightarrow Y_5 \rightarrow Y_6. \quad (4f)$$

Unfortunately, by unblocking the colliding paths containing  $Y_1 \rightarrow Y_3 \leftarrow Y_2$ , it *creates* the

new  $d$ -connecting paths

$$Y_4 \leftarrow Y_1 - Y_2 \rightarrow Y_5 \rightarrow Y_6, \quad (5a)$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 - Y_2 \rightarrow Y_5 \rightarrow Y_6, \quad (5b)$$

$$Y_4 \leftarrow Y_1 - Y_2 \leftarrow C_{2,5} \rightarrow Y_5 \rightarrow Y_6, \quad (5c)$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 - Y_2 \leftarrow C_{2,5} \rightarrow Y_5 \rightarrow Y_6. \quad (5d)$$

The paths in (5) use an *undirected edge* between two variables to indicate that they are  $d$ -connected only after conditioning on their common descendant.

Path (5a) presents a simple case of unblocking a collider.  $Y_1$  is a parent of  $Y_4$ , and  $Y_2$  is an ancestor of  $Y_6$ . Thus, once we conditionally confound  $Y_1$  and  $Y_2$ , the causal flow from these nodes creates an additional  $d$ -connecting path between  $Y_4$  and  $Y_6$ . Path (5d) is instructive; contrary to Wright's (1934; 1968) rules, tracing this path to induce a covariance between  $Y_4$  and  $Y_6$  is legitimate despite having to go backward after already going forward. The justification of this is that after we condition on the common descendant of two causal lineages, each ancestor in one lineage will find itself  $d$ -connected with every ancestor in the other lineage. This must be true because the length of a directed path is a feature of human knowledge rather than external reality; therefore it must be possible to go from  $C_{1,4}$  to  $C_{2,5}$  regardless of whether any mediators along the way to the unblocked collider  $Y_3$  are known. Thus, we can trace backward from  $Y_4$  to the unobserved confounder  $C_{1,4}$ ; this confounder is connected to  $C_{2,5}$ , from which we can proceed forward through  $Y_5$  to arrive at  $Y_6$ .

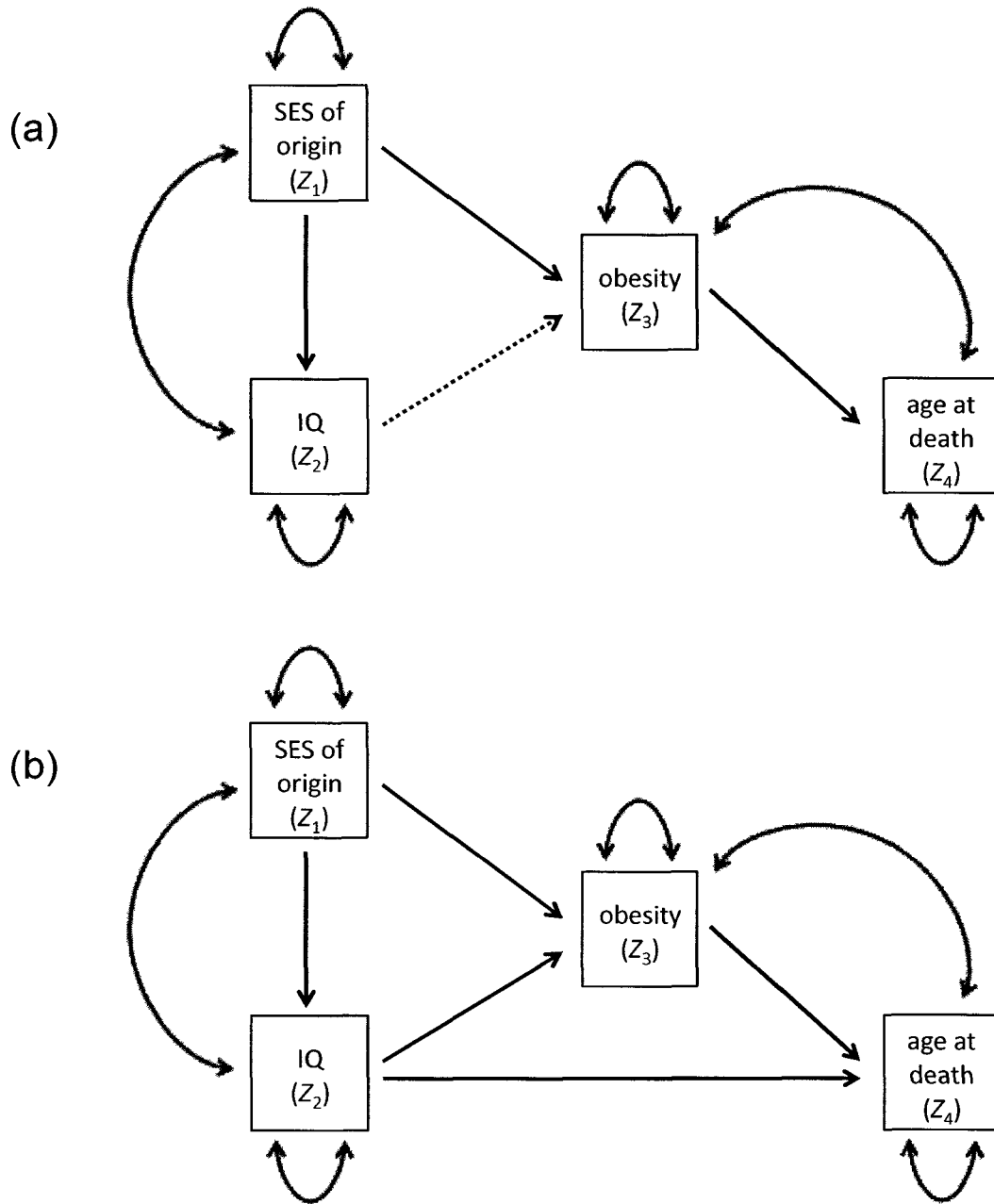
To summarize, the collision at  $Y_3$  normally impedes any causal flow through the paths in (5). Conditioning on  $Y_3$  unblocks the collision and allows the paths to  $d$ -connect  $Y_4$  and  $Y_6$

in the offspring. That is, among households *observed* to have the same SES, the covariation among the causes of SES is altered, probably becoming more negative. Thus, whenever we have two such causes of SES, each also affecting a different member of the pair  $\{Y_4, Y_6\}$ , they suppress the estimated magnitude of any  $Y_4 \rightarrow Y_6$  causal effect. Conditioning on any member of  $\{Y_1, Y_2, Y_5\}$ , in addition to  $Y_3$ , will restore these colliding paths to their original  $d$ -separated status. If we have not measured any of these variables, however, at best we can hope that the statistical control for parental SES removes more bias than it introduces.

The next example shows that whether we should condition on a particular variable can depend on whether we want to test a sharp null hypothesis or to estimate the size of a causal effect. In recent years several studies have shown that personality traits are associated with longevity (Friedman et al., 1993; Gottfredson & Deary, 2004; Batty et al., 2009; Gallacher et al., 2009). Much as in Figure 14, these studies attempt to control for possible confounders and also to determine how much of any effect is attributable to mediators such as obesity, smoking, and attained SES. Figure 14 incorporates the SEM custom of using a bidirectional arc that begins and ends at the same node to represent the residual disturbing causes. Explicit representation of the disturbances can greatly assist our understanding of a model, reminding us that each variable has other causes not depicted as nodes.

In this example we refrain from any parametric assumptions. What this means is that in the place of regression coefficients, we consider quantities of the form  $E[(Y | do(x')) - (Y | see(x))]$  for selected values of  $\{x', x\}$ . Nonparametric estimation of conditional probability density functions is of course not a simple matter. However, it is nevertheless useful to neglect practical considerations and concentrate on the theoretical points, in order to realize that SEM is not intrinsically tied to linear causal relations. For the sake of argument,

Figure 14: A DAG representing a causal model of intelligence and mortality.



we initially assume that  $Z_2 \rightarrow Z_4$  is absent (Figure 14a). That is, intelligence only impacts mortality indirectly through obesity; perhaps smarter people are better informed about the dangers of being overweight.

In his responses to skeptical readers of *Causality*'s first edition, Pearl (2009, pp. 340-341) addresses an epidemiologist who claims that tradition allows for conditioning on both  $Z_1$  and  $Z_4$  to evaluate the effect  $Z_2 \rightarrow Z_3$ . This is permissible and in fact desirable—if we simply wish to test the null hypothesis that IQ has no effect on obesity. Conditioning on  $Z_4$  alone is actually a step backward from this goal because  $Z_4$  is a descendant of the collider  $Z_3$ . Let us pause to consider why.

In the subgraph representing the null hypothesis, there is no directed edge  $Z_2 \rightarrow Z_3$ , which means that the only node sending an edge to  $Z_3$  is  $Z_1$ . What, then, makes  $Z_3$  a collider? Recall that the disturbance of  $Z_3$ —which we now call  $E_3$ —represents unmeasured causes of obesity. *These* causes are now spuriously associated with  $Z_1$  because of the conditioning on their common effect. As the unmeasured causes of obesity fluctuate, so does childhood SES, which in turn passes causal flow to IQ. Fortunately, conditioning on  $Z_1$  blocks the newly opened path  $Z_2 \leftarrow Z_1 - E_3 \rightarrow Z_3$  and also the original confounding paths. Thus, if there remains an association between IQ and obesity at some  $\{z_1, z_4\}$ , we must reject the null hypothesis that the subgraph lacking the edge  $Z_2 \rightarrow Z_3$  is correctly drawn. The most natural reason for the rejection of the hypothesis would be that there is in fact such an edge.

Once we accept the alternative hypothesis that  $Z_2 \rightarrow Z_3$  is present, the set  $\{Z_1, Z_4\}$  is no longer an admissible set of covariates for *estimating* the  $Z_2 \rightarrow Z_3$  effect. Mortality is now a descendant of IQ ( $Z_2 \rightarrow Z_3 \rightarrow Z_4$ ), which means that conditioning on mortality *d-*



connects IQ and the unmeasured causes of obesity by opening the path  $Z_2 - E_3 \rightarrow Z_3$ . This unblocked path cannot be  $d$ -separated by any node in Figure 14. If we want to estimate the effect of IQ on obesity, we must condition only on SES. But  $Z_4$  is still a desirable covariate for testing the null hypothesis (*IQ does not affect obesity*) because the continued conditional independence of  $Z_2$  and  $Z_3$  after adding  $Z_4$  to the covariate set is an additional constraint on that hypothesis.

Let us reintroduce the directed edge  $Z_2 \rightarrow Z_4$  (Figure 14b). That is, IQ now affects mortality through mechanisms other than obesity. Because mortality is now a descendant of IQ, it is no longer an admissible covariate for any evaluation of  $Z_2 \rightarrow Z_3$ . We can still estimate the effect of IQ on obesity by conditioning on SES. But can we also estimate the direct effect of IQ on mortality?  $\{Z_1, Z_3\}$  may be a tempting set of covariates for this purpose. Again, however, IQ and the disturbance of obesity collide at obesity itself. Conditioning on obesity will thus  $d$ -connect IQ with the unmeasured causes of obesity, including whatever unmeasured common causes obesity shares with mortality. This unblocks the path  $Z_2 - C_{3,4} \rightarrow Z_4$ . It turns out that we cannot estimate the desired direct effect without imposing some parametric assumptions. Given only the graphical assumptions regarding the connectivity of the nodes, we can only estimate the *total* effect of IQ on mortality (by conditioning on  $Z_1$ ).

The point of these exercises is not to argue for any particular model or claimed empirical finding. It is rather to dissuade readers from the belief that a conditioning technique such as multiple regression is a more innocuous method than full-fledged SEM. The very opposite is true. Since multiple regression is a linear model of what we would find given certain *observations*, it can only tell us what we would find given certain *actions* under special conditions. By implicitly making assumptions that a graphically rendered model always makes explicit,

multiple regression is actually by far the more equivocal method. The lesson is clear: *when making inferences from observational data, we should always present a (graphical) structural equation model representing our causal theory so that its critical assumptions can be criticized and defended.* In fact, one might hope that disagreements over the interpretation of observational data will often reduce to disagreements over how to connect each pair of nodes. Once the nature of the disagreement becomes this explicit, both sides should find it easier to decide whether the existing data rule out any contending hypothesis and also whether any additional data can be collected to narrow the divide between them.

That said, in cases where the linearity approximation is reasonable, there is still an important role for regression in causal analysis. At the very least, we may continue to encounter the naive use of multiple regression in the literature, and criteria for whether a partial regression coefficient gives an unbiased estimate of the desired causal effect are useful in judging such analyses. Furthermore, if the causal model is not globally identified, we might be forced to use regression to estimate those causal effects that are locally identified. In fact, for many reasons, some of which are discussed in Part 3, the local identification of effects is far more informative than the bare fact of global identification.

To identify any partial effect in a linear model, as defined by a selected set of direct or indirect paths from  $X$  to  $Y$ , we must find a set  $\mathbb{S}$  of measured variables that contains no descendant of  $Y$  and  $d$ -separates all non-selected paths between  $X$  and  $Y$ . The partial effect will then equal the regression coefficient of  $X$  in the multiple regression of  $Y$  on  $X \cup \mathbb{S}$  (Pearl, 1998; Spirtes et al., 1998).

Whenever a report presents a partial regression coefficient as an estimate of a causal effect, it may be useful to construct plausible DAGs (structural equation models) and determine

which of these satisfy the conditions of the theorem.

The approximation of linear causal relations will sometimes be untenable. This would possibly be the case in Figure 14 if mortality were measured dichotomously at one time point. In these situations methods are available for other functional forms, many of which have been developed by statisticians and econometricians (Lee, 2007; Wooldridge, 2010). A promising direction for sufficiently low-dimensional problems is to dispense with parametric assumptions and work directly with the sample estimate of the joint probability distribution, using an ingenious calculus devised by Pearl for his *do* operator. Of course, the absence of parametric assumptions weakens results on whether an unbiased estimate of an effect is identified. For example, the direct effect  $Z_2 \rightarrow Z_4$  in Figure 14 is unidentified in general, but becomes identified in a linear model because the identified total effect of  $Z_2$  on  $Z_4$  is then merely the sum of products given by the path-tracing rules. We identify coefficients in other terms of this sum and then subtract the result from the total effect to obtain the edge coefficient of interest. Interested readers should carefully study Pearl's (2009) treatment of this matter. Nonparametric estimation needs extremely large samples, but many studies of personality and health appear to satisfy this requirement.

### 3.1.2 The Value of Randomization

Imagining the experiments implied by each directed edge can sharpen our justifications for including and omitting arcs. Of course, the best way to ensure the feasibility of some experiment is to actually perform it. Moreover, the graphical theory of causality justifies *randomly* assigning levels of the putative causal variable whenever this is feasible. Over much resistance by seasoned experimenters, Ronald Fisher advocated randomization for the pre-

cise purpose of distinguishing causation from confounding (Box, 1978). This is one of the stark contrasts between Fisher and his nemesis Karl Pearson; Fisher did not regard causality as a meaningless concept. Although his argument from “the lady tasting tea” is characteristically difficult, I believe that we can rephrase Fisher’s (1966) defense of randomization in graphical terminology as follows. By assigning subjects to different levels of a putative cause according to a random mechanism, we are *d*-separating the variable from all of its ancestors in the causal graph containing it (Figure 12b). Since a coin flip is by design unassociated with any macroscopic variable, it shields the putative cause from any confounders lurking among its ancestors or the experimenter’s whims. If the actual *implementation* of the manipulation matches the ideal of surgical isolation represented by the *do* operator, we can assess the resulting association between putative cause and effect for significance against an exact sampling distribution.

The deep insight in this rationale for randomization is characteristic of Fisher’s writings on causality. We will return to these writings when we consider causality in the context of genetic research. For now I point out that the graphical theory neatly unites the innovations in causal inference developed by both Fisher and his great rival in population genetics, Sewall Wright. Path analysis, as practiced ingeniously by Wright (1921; 1931; 1969), is an obvious intellectual forerunner of the graphical theory. The characterization of Fisherian randomization as edge deletion shows the great conceptual generality of this theory.

The rarity of such powerful natural experiments may seem to leave randomization as a peripheral concept to personality research. In the spirit of Pearl’s call to “causation without manipulation,” however, we should recognize that randomization, fixing the values of potential confounders, and even conditioning on colliders are not the prerogatives of human

scientists. Nature herself engages in these activities; Part 4 will have much more to say about this.

### 3.2 The Nature of Psychometric Common Factors

Since two personality scales with no items in common can be functionally interchangeable, whatever is measured by any single scale must be somehow generalizable beyond its specific items. In addition, two interchangeable scales will always show a correlation less than unity, so we must suppose that the scales measure a third quantity with some degree of error. Personality researchers have often looked to *common factors* as the generalizable quantities that any particular scale imperfectly measures.

The mathematical conception of a common factor is perhaps clear enough. But any mathematical model must be understood as providing an analogy to some external reality, and thus the question arises: what exactly in the real world does a common factor represent? This issue has provoked intense and recurrent debate among psychometricians. Mulaik (2005) ably reviews certain aspects of the controversies; noteworthy recent contributions include Borsboom, Mellenbergh, and van Heerden (2003), Bartholomew (2004), and Ashton and Lee (2005). No writer seems to have convincingly settled the issue in the compass of a single article (or book), and I will not try to be the first. But the statement of some position, however brief and debatable, is called for here in order to move on with our attempts to employ psychometric common factors in causal explanations. In what follows I rely heavily on McDonald (1996; 2003).

Factor-analytic models treat measured variables, such as the different items or subscales in a personality battery, as indicators of unmeasured quantitative variables called *factors*

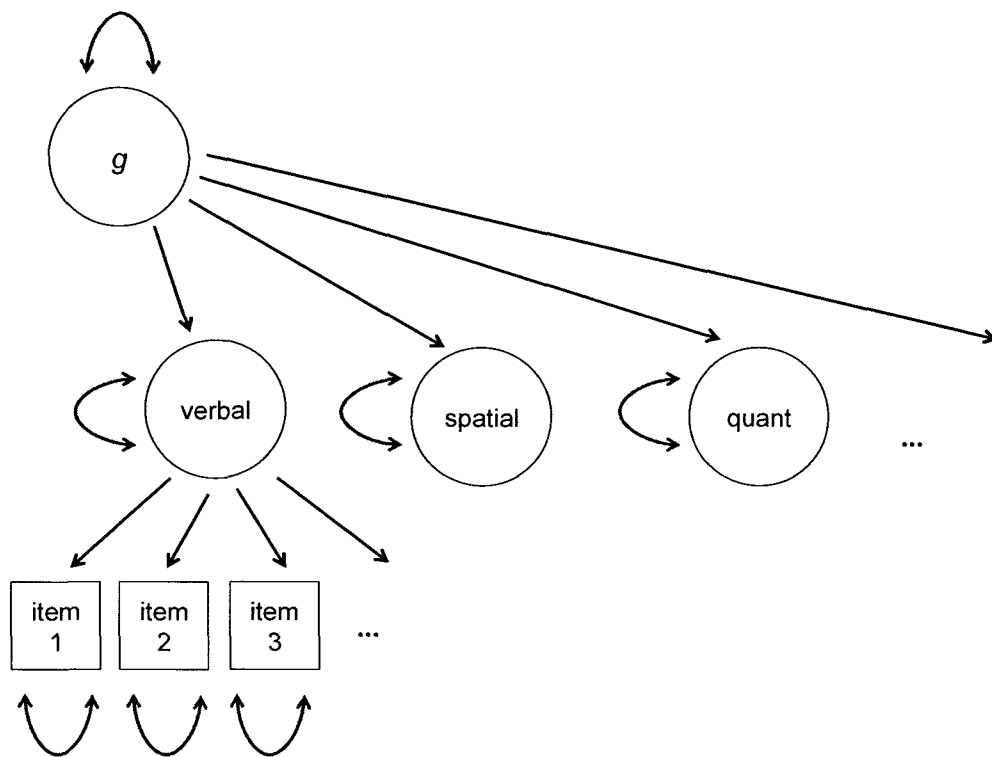
(Thomson, 1951; McDonald, 1985; Mulaik, 2010). The term *factor* here has a narrower meaning than when used as a rough synonym for *cause* or *variable*. If the scores on a scale could be regressed on the unobserved factor scores, each regression coefficient would represent the quality of the scale as a measure of the corresponding factor. The regression coefficients in this model are called *factor loadings*.

Figure 15, a graphical depiction of a schematic factor model, introduces another SEM diagramming convention: the depiction of nodes standing for common factors (sometimes called *latent variables*) as circles instead of rectangles. Despite the similarity of Figure 15 to those examined in Part 1, I maintain that the coefficients (factor loadings) attached to the directed edges here should *not* be interpreted as representing the magnitudes of causal effects. A factor model is not necessarily a causal model.

Suppose that we use the dimensions and weights of various body parts as indicators of a common factor called *body size*. This is a common conceit in didactic accounts of factor analysis, most recently taken up by Bartholomew (2004). Now consider the proposal that body size is the unobserved common cause of height, weight, and so forth. To most of us, at least, the notion that size causes height will seem very close to circular. In one terminology the relation between the concepts of size and height seems much too “analytic” to permit construing it as a causal one. Indeed, causal inference is a problem precisely because causation is a “synthetic” relation. It turns out to be true that the bacterium *Vibrio cholerae* causes the disease it is named after, but logically it could have been otherwise and therefore this relation had to be empirically discovered.

Body size is not the common cause of those variables that measure it, but rather is their common abstractive property. Furthermore, the large loading of a given indicator on size

Figure 15: Each set of items at the lowest level defines a factor. The first-order factors are in turn measures of a higher-order factor called  $g$  (general intelligence).



does not imply that there is some unobserved variable (but observable in principle), which, when severed from its ancestors and adjusted upward by one unit, will yield an increase in the value of the indicator equal to the loading. A large loading simply means that there is a high degree of conceptual overlap between the (unobservable in principle) abstract property and the (observable) indicator. Height is not exactly the same as body size, but it is a good proxy for it. We might say that height makes for a passable “Size Quotient.”

I believe that the analogous assertion holds for psychometric common factors. Consider the relation between extraversion and whether the respondent likes to meet new people. An indubitable meaning of the statement *A likes to meet new people because he is extraverted* is that the respondent’s behavior in this instance has an “intensity” that is typical of his behavior in a wide class of semantically related instances: whether he likes to attend parties, whether he goes out of his way to greet people, whether he feels comfortable speaking in front of groups, and so on. But if we construe the relation between extraversion and meeting new people as a causal one, we are essentially saying that an abstraction of the respondent’s behavior across a class of instances causes his behavior in a particular instance: being extraverted causes a behavior typical of an extravert. That is, unlike the relation between *Vibrio cholerae* and cholera, the relation between extraversion and meeting new people fails to offer a means of defining the putative cause and effect independently of one another.

Someone determined to rescue the notion of a common factor as a common cause of its indicators might claim that general intelligence (*g*), extraversion, and other *psychometric traits* do not in fact correspond to the *folk-psychological traits* bearing these names. According to this argument, just as the physical construct of gravity bears only a metaphorical resemblance to the natural-language concept (*weight* or *seriousness*), the Big Five/Six trait



of extraversion bears a resemblance of a similar kind to the natural-language concept while in fact having a distinguishable intension (presumably some neural attribute). Perhaps the simplest objection to this argument is that it is out of harmony with the actual behavior of personality researchers and other applied psychometricians. When psychometricians want to increase the reliability of a scale, they add more indicators of the “same kind”—more items eliciting either right or wrong answers, more items inquiring about religious proclivities. This is rather telling evidence that users of factor analysis do not treat common factors as common causes. It would be a rather curious restriction on the effects of the same cause that they must all share some namable psychological-semantic property. Consider Newton’s striking unification of celestial and terrestrial mechanics. What in our *a priori* semantics could possibly allow us to construe the fall of an apple, the oscillation of the tides, and the orbits of heavenly bodies as belonging to the “same kind?”

Perhaps this is enough to convince the reader that interpreting the factor underlying a set of indicators as an abstraction of the set’s semantic commonality is at least as convincing as a causal interpretation. But what of the factor’s relations to external variables? Can *these* said to be causal? Although there can be no doubt that an abstraction such as body size might be a useful predictor, can body size really be said to *cause* anything? The answer to this question seems to be yes—if transforming someone’s body so that he must be assigned a different size factor score is a conceptually permissible manipulation. The causal claim *A won the fight because he is bigger than B* then amounts to the following: if we could have fixed A’s factor score to a sufficiently low value—perhaps by transplanting A’s mind to a much smaller body—then A would not have prevailed over B. Models in which other variables appear as causes of a common factor may also prove to be very useful approximations;

McDonald (1996) provides the example of alcohol temporarily increasing extraversion.

In fact, if one accepts that factor analysis by itself is not a tool for the discovery of causes, causality only enters the picture when we consider relations with external variables. If we could complete a causal chain like the one depicted in Figure 11, what traits would we most want to insert in the place of the node labeled *trait variation*? An evolutionarily oriented psychologist might choose those traits figuring in important theoretical accounts of human evolution. Ashton and Lee (2001) take this line in defending their HEXACO model of personality variation. They have chosen a basis where three of the six axes are defined by behaviors of critical importance in evolutionary theories of human cooperation: Emotionality (responding to feelings of kinship and solidarity), Agreeableness (initiating exchanges, forgiving defectors), and Honesty (never defecting first, reciprocating favors). Psychologists studying other domains of individual differences might adopt this approach. Instead of attempting to find a periodic table of traits, we should try to ensure that our instruments measure traits whose causes and consequences are worth understanding. Supplying such rationales may seem to assume the presence of the links in Figure 11 that we are trying to establish, but surely this circularity is not a vicious one.

To summarize, factor analysis is a tool for refining the measurement of abstractive traits that are hypothesized to exist in advance of any data analysis. Such a trait is not a common cause of the indicators used to measure it, but this does not mean that the trait is not real. Dismissing intelligence, extraversion, liberalism, religiosity, and the like as mathematical fictions would decimate our causal understanding of social reality. If these traits are at all fictional, they are fictions of folk psychology. The adoption of psychometric methodology implies a commitment to the view that the insertion of traits, moods, and other *intervening*

*variables* of folk psychology between brain and behavior has proven fruitful and will continue to be necessary (MacCorquodale & Meehl, 1948).

We now have a perhaps complete taxonomy of reasons for an observed correlation between variables  $X$  and  $Y$ :

1.  $X$  is a cause of  $Y$  (or vice versa).
2.  $X$  and  $Y$  are both effects of a common cause.
3.  $X$  and  $Y$  are both causes of a collider that has been conditioned on.
4.  $X$  and  $Y$  are both measures of an abstractive property.

These reasons may not be mutually exclusive for a given  $X$  and  $Y$ . The last reason can never hold in the absence of at least one other.

In Part 3 we resume our applications of the graphical approach, realizing now that our aim is to verify the adequacy of the approximation entailed by employing common factors in causal explanations.

### **3.3 The DAG As a Source of Severe Empirical Tests in Structural Equation Modeling**

In Part 1 we looked at SEM from a certain perspective. We took some of the nodes in Figure 13 to be unmeasured, ruling out any chance to estimate the parameters associated with them. Even if these nodes were measured, the bidirectional arcs would defeat the sufficient condition for the global identifiability of a linear model (Brito & Pearl, 2002). Figure 14b does not meet the well-known necessary condition for the global identifiability of a linear model, and in any case we considered Figure 14 in a nonparametric setting. We asked the

following question: taking the depicted system of causal relations more or less for granted, what is a necessary and sufficient set of covariates for locally identifying an unbiased estimate of a causal effect? We decided that a proper answer to the question cannot neglect the causal DAGs (generally nonlinear structural equation models) containing the putative cause and effect. But before arriving at that conclusion, the fact that we were undertaking SEM may not have been readily apparent.

Here we consider SEM from a perspective more closely aligned with the traditional one. For this reason the contrast between conventional SEM practice and Pearl's graphical approach should be evident throughout. Given a causal DAG where all depicted nodes have been measured and global identification obtains, we place the task of estimation somewhat in the background and ask the following question: what assurance do we have that the causal model, as drawn, reflects reality to an acceptable degree of approximation? The orthodox response to this vital question emphasizes the simultaneous analysis of all measured variables and global goodness-of-fit. But because this approach by itself does not foreclose certain logical absurdities, it should at the very least be supplemented by the approach advocated in this article.

Taken at face value, the orthodox view accepts the plausibility of the model

$$\Omega = \{\text{barometer readings cause rain}\} \cup \{\text{the average age in Los Angeles is higher than three}\}.$$

When confronted with actual measurements,  $\Omega$  will fit the data extremely well and escape falsification. The problem is that a strong correlation between certain barometer readings

and rain, combined with the average age in Los Angeles being well over three, tells us nothing about whether barometers cause rain. We must therefore insist that the tested component of  $\Omega$  (*the average age is higher than three*) bear a logical relation to what  $\Omega$  claims (*the correlation between certain barometer readings and rain means that barometers cause rain*).

Combining the factor and causal models in one graph and then estimating all parameters in one global fit is a prime example of conjoining causal claims to components that are essentially independent of them. A common procedure among personality researchers is to fit a hybrid factor-causal model and apply a rule of thumb to a scalar index of model misfit such as the root mean square error of approximation (RMSEA). But in cases where the factor model fits extremely well (which it typically will in well-motivated applications), the causal model can fit poorly without the misfit being reflected in the scalar index. This is not a fanciful objection. McDonald and Ho (2002) reexamined 14 studies basing their conclusions only on global fit indices and found that in nine of these the causal model fit poorly by any reasonable standard. McDonald (2010) provides a detailed empirical case study of how a global fit can lead to wholly erroneous conclusions.

It seems fruitful, then, to effect a clean divorce between measurement and causation through Anderson and Gerbing's (1988) two-step procedure: (1) estimate and test the adequacy of only the factor model, freely estimating the covariances among the factors and any non-factor variables, and then, if this step succeeds, (2) fit the causal model to the resulting covariances. Even this procedure, however, suffers from potential blurring of misfit. If there is a substantial local deviation of the data from what the model allows, adjustments in fitting other parts of the model may compensate for the discrepancy, resulting in a scatter of small and innocent-seeming elements in the residual covariance matrix.

What we therefore require is a method that permits *local* tests of whatever predictions are entailed by a causal model. Here is where Pearl's principle of *d*-separation becomes applicable. Recall that two variables are conditionally independent given the covariates in their *d*-separating set. Conditional independence implies a vanishing partial covariance between the two variables. We can thus simply list the vanishing partial covariances implied by a causal model and examine each one for its numerical closeness to the point prediction of zero (Shipley, 2000). An additional and perhaps surprising advantage of this approach is that it is valid for arbitrary functional forms of the causal relations and arbitrary distributions of the disturbances.

To illustrate this elegant procedure, I reanalyze a dataset presented by Deary, Batty, and Gale (2008). Based on a sample of 3,412 males and 3,658 females, the authors conclude that a higher level of *g* (measured at age 11) is both a direct and indirect cause of more liberal social attitudes (measured at age 30). Figure 16 depicts their preferred model.<sup>3</sup>

We first note that any given DAG entails many vanishing partial covariances, not all of which are independent of the others. Fortunately, there exists a *basis set* of independent partial covariances; if all of the partial covariances in this basis set equal zero, every partial covariance predicted to equal zero will in fact do so (Pearl & Verma, 1987). In this class of models, the basis set consists of the partial covariances  $\sigma_{i,j-\{\text{parents of } j\}}$  for all *i* preceding *j*

---

<sup>3</sup>Deary and colleagues allow a directed edge to connect a causally prior observed variable with one of their subscales measuring liberalism. Allowing such edges is problematic because they may prevent the common factor of the subscales from satisfying a property that psychometricians call the *principle of local independence* (Lord & Novick, 1968; McDonald, 1981). Although I believe that the common factors of a behavior domain must (approximately) satisfy this principle, this is perhaps debatable. To avoid discussing the merits of this issue, we will retain only one of the subscales used by Deary and colleagues. Arbitrarily, we choose the subscale called *antiracism*. In the factor model, we fix the standardized loading of the subscale on its common factor to the square root of its reliability.

Figure 16: A DAG representing a causal model of the variables studied by Deary et al. (2008).

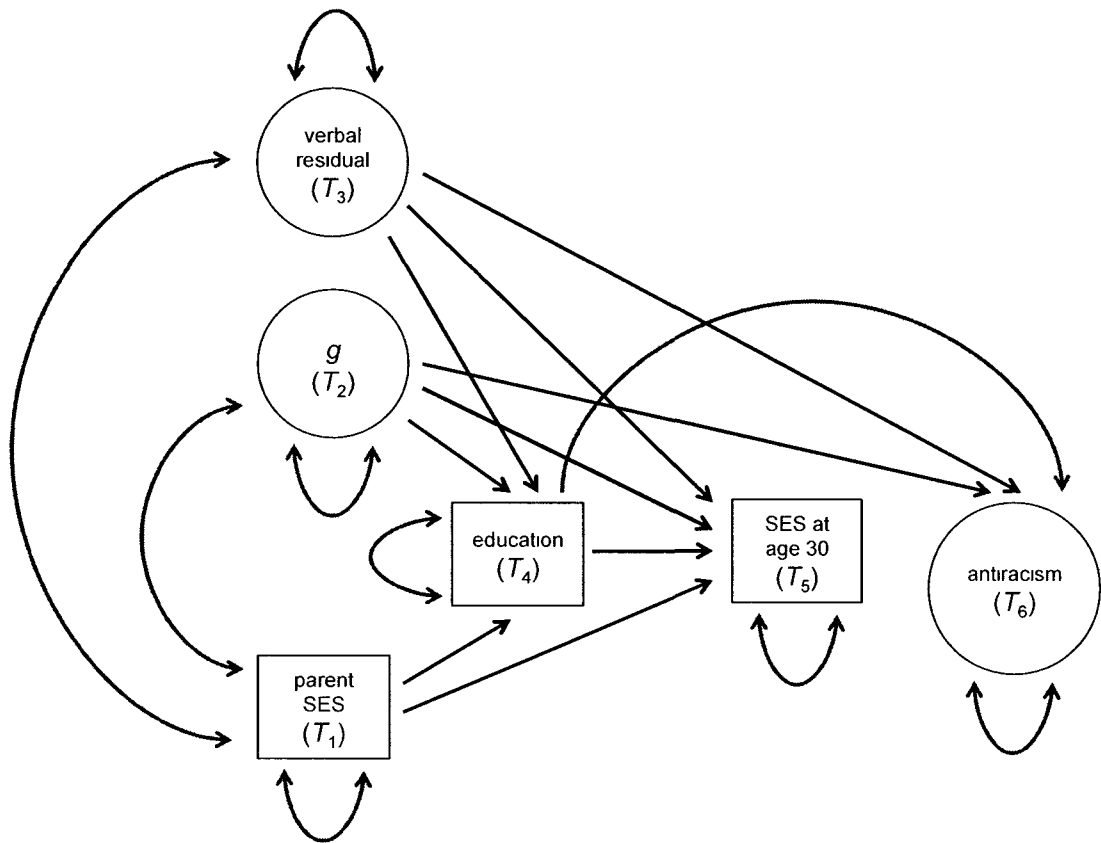


Table 12:  $d$ -separation tests of the causal model in Figure 16

$d$ -separable nodes	$r_{i,j \{ \text{parents of } j \}} (95\% \text{ CI})$		$p$ -value	
	Male	Female	Male	Female
$T_1, T_6$	-.003 (-.036, .030)	-.030 (-.062, .001)	.87	.07
$T_5, T_6$	.058 (.025, .091)	-.028 (-.059, .004)	.0007	.10

*Note.*  $r_{i,j \{ \text{parents of } j \}}$  stands for the correlation between  $i$  and  $j$  after partialing out the parents of  $j$ . The  $p$ -values in each column can be combined by Fisher's method to provide an overall test of the model for males ( $\chi^2_4 = 14.9, p < .005$ ) and females ( $\chi^2_4 = 10.1, p < .05$ ).

where the two variables are not connected by a directed edge. This characterization of the basis set is quite intuitive. If node  $i$  is not among the parents of node  $j$ , conditioning on these parents shields  $j$  from all  $d$ -connecting paths to  $i$ .<sup>4</sup>

In our own example, we proceed by finding each pair  $\{T_i, T_j\}$  not connected by a directed edge. There are three such pairs:  $\{T_1, T_2\}$ ,  $\{T_1, T_6\}$ , and  $\{T_5, T_6\}$ . Since the first of these pairs consists of definitionally orthogonal common factors, there are only two point predictions: given appropriate sets of covariates, the partial covariances of  $\{T_1, T_6\}$  and  $\{T_5, T_6\}$  are equal to zero. The substantive import of these predictions is that neither parental SES nor attained SES at age 30 has a direct effect on racial tolerance. These are fairly remarkable claims. One might have thought that moving up or down the occupational hierarchy might at least affect exposure to individuals of different races, leading in turn to changes in antiracism.

<sup>4</sup>An important caveat is that this approach exhausts the testable implications of a given DAG if the model is *exogenous*—that is, if the only variables connected by bidirectional arcs are those whose causes are not specified within the model. *Endogenous* models may imply point predictions that do not take the form of vanishing partial covariances. Critically, it is not known whether there is a general method for finding a basis set implying all of the point predictions entailed by an endogenous model. Although McDonald (2002, 2004) provides methods for endogenous models, these do not seem straightforward to apply. This is an area requiring further research. In the meantime the program Commentator does supply all point predictions entailed by an endogenous model, regardless of whether they can be reduced to a basis set.



Table 12 presents the results of the  $d$ -separation tests. The confidence intervals are rather wide, which shows that 3,000 participants does not approach the point of diminishing returns in the graphical approach to SEM. Despite the ambiguities we try to interpret the results that we have.

Since the overall model is rejected in both sexes, we are forced to a judgment of whether the numerical discrepancies are still small enough to earn the model “money in the bank.” The  $T_5$ – $T_6$  partial correlation in males is the most discrepant. The sign of this partial correlation in females has the opposite sign, however, suggesting that the source of the discrepancy is small or unsystematic. The  $T_1$ – $T_6$  partial correlations, particularly in males, do seem to be vanishing.

The essential claim encoded in Figure 16 is that social status, at all parts of the lifespan up till age 30, has no direct effect on antiracism. The  $d$ -separation tests provide much stronger support for this claim than most theory-testing methodologies in observational behavioral science. First, the conclusions do not depend on assumptions regarding the functional forms of the causal relations or the distributions of the disturbances. Second, the support for the theoretical claims draws on the closeness of observations to risky point predictions (Meehl, 1990). Third, the locality of the testing means that we can diagnose *where* the model has gone astray in cases where the data miss the predictions. Suppose that in our judgment the partial correlation between  $T_5$  and  $T_6$  in males is too large to support the model. We must then somehow ensure that these two nodes are  $d$ -connected even after conditioning on  $\{T_2, T_3, T_4\}$ . Note that insertion of the directed edge  $T_5 \rightarrow T_6$  will also  $d$ -connect  $T_1$  and  $T_6$ . If we are satisfied that these latter two nodes are truly  $d$ -separated by  $\{T_2, T_3, T_4\}$ , then instead of  $T_5 \rightarrow T_6$  we might prefer to insert the reversed edge  $T_6 \rightarrow T_5$ , the bidirec-

Table 13: Parameter estimates of the causal model in Figure 16

Parameter	Unstandardized (SE)		Standardized		Deary et al. (2008)	
	Male	Female	Male	Female	Male	Female
$\beta_{4,1}$	.163 (.019)	.203 (.017)	.14	.18	.14	.18
$\beta_{4,2}$	.518 (.018)	.504 (.018)	.45	.42	.42	.42
$\beta_{4,3}$	.032 (.020)	.128 (.024)	.02	.08		
$\beta_{5,1}$	.136 (.016)	.075 (.015)	.13	.08	.13	.08
$\beta_{5,2}$	.239 (.017)	.212 (.017)	.24	.21	.23	.19
$\beta_{5,3}$	.049 (.017)	-.006 (.020)	.04	-.00		
$\beta_{5,4}$	.260 (.014)	.253 (.014)	.30	.31	.31	.31
$\beta_{6,2}$	.174 (.017)	.174 (.015)	.19	.21	.18	.18
$\beta_{6,3}$	.076 (.017)	.174 (.018)	.07	.15		
$\beta_{6,4}$	.105 (.015)	.092 (.012)	.13	.13	.09	.09
$\psi_{1,1}$	1.184 (.029)	1.206 (.028)	1	1	1	1
$\psi_{2,2}$	1.290 (.031)	1.110 (.026)	1	1	1	1
$\psi_{3,3}$	.960 (.023)	.565 (.013)	1	1		
$\psi_{4,4}$	1.268 (.031)	1.158 (.027)	.78	.78	.80	.79
$\psi_{5,5}$	.908 (.022)	.834 (.019)	.86	.88	.83	.86
$\psi_{6,6}$	.997 (.024)	.674 (.016)	.95	.92		
$\psi_{1,2}$	.400 (.022)	.396 (.020)	.32	.34	.38	.40
$\psi_{1,3}$	.126 (.017)	.092 (.013)	.12	.11		

Note.  $\beta_{i,j}$  stands for the direct causal effect of  $T_j$  on  $T_i$ .  $\psi_{i,j}$  stands for the residual covariance of variables  $T_i$  and  $T_j$ . The estimates of the underlying factor model have been omitted for brevity.

tional arc  $T_5 \leftrightarrow T_6$ , or both. That is, in males at least, if there is no confounding of  $T_5$  and  $T_6$ , then SES at age 30 does not directly affect antiracism but rather the other way around. Upon reflection this hypothesis is perhaps a natural one; nowadays being a frank racist may hurt one's career prospects.

Although the *absence* of directed edges from social status to antiracism is no doubt an interesting finding, the primary issue in this study is the *presence* of a directed edge from  $g$  to antiracism. Table 13 gives the maximum-likelihood estimates of the causal parameters

in the linear model. The RMSEA in males is .030; in females, .016. Among the variables prior to antiracism,  $g$  is estimated to have the largest standardized direct effect ( $\beta_{6,2}^* \approx .20$ ). But now we face our key question: what has our graphical analysis revealed so far about the trustworthiness of this estimate? If the model survives the risk posed by its predicted vanishing partial covariances, how much should our ensuing confidence extend to parts of the model other than the  $d$ -separable nodes? Pearl (2004) provides a general discussion of the relationship between the robustness of important effect estimates and the validity of assumptions regarding absent edges. We can work out our own special case to bring out the main ideas.

Readers familiar with the notion of *covariance equivalence* will know that for any given causal model there may exist several distinct models that produce exactly the same fit to the covariance matrix. A trivial example is the chain  $X \rightarrow Y \rightarrow Z$ , which is covariance equivalent to the reversed chain  $Z \rightarrow Y \rightarrow X$  and the common-cause model  $X \leftarrow Y \rightarrow Z$ . Each of these three models implies the same vanishing partial covariance:  $\sigma_{XZ.Y}$ . In a certain class of models, it is generally true that two causal models are covariance equivalent if and only if they entail the same set of vanishing partial covariances.<sup>5</sup> This graphical perspective is valuable because it provides an intuitive means of ascertaining whether a deprecated model may be covariance equivalent to the preferred one. If two nodes that are  $d$ -separable in the preferred model are no longer  $d$ -separable after some alteration, then the new model is not covariance equivalent to the preferred model. Pearl (2009) provides an ingenious elaboration of this insight.

---

<sup>5</sup>When the class of competitor models includes endogenous models, entailing the same vanishing partial covariances is only a necessary condition because endogenous models may impose constraints that do not take the form of vanishing partial covariances.

Temporal considerations weigh against most of the edge reversals otherwise permitted in Figure 16. The assumption that is most critical to the validity of  $\hat{\beta}_{6,2}$  is thus the absence of a bidirectional arc between  $T_2$  and  $T_6$ . If interchanging  $T_2 \rightarrow T_6$  and  $T_2 \leftrightarrow T_6$  (or simply adding  $T_2 \leftrightarrow T_6$ ) preserves all vanishing partial covariances, we can place no confidence in our obtained  $\hat{\beta}_{6,2}$ ; the relation between  $g$  and antiracism may instead be attributable in its entirety to confounding. The  $d$ -separability of  $\{T_1, T_6\}$  and  $\{T_5, T_6\}$ , however, forbids the presence of  $T_2 \leftrightarrow T_6$ . If there is such a bidirectional arc, then conditioning on  $T_2$  opens the colliding path  $T_1 \leftarrow C_{1,2} - C_{2,6} \rightarrow T_6$ , which cannot be blocked by any measured variable. In fact, if there were a confounder of  $T_2$  and  $T_6$  inducing a correlation of .20 between these two variables, conditioning on  $T_2$  would induce a correlation of roughly  $-.07$  between  $T_1$  and  $T_6$ . In summary,  $T_2 \rightarrow T_6$  and  $T_2 \leftrightarrow T_6$  do not predict the same vanishing partial covariances, and thus the near-zero values of the partial covariances predicted to vanish specifically under the direct effect  $T_2 \rightarrow T_6$  provide evidence against confounding of the form  $T_2 \leftrightarrow T_6$ .

A similar argument shows that our  $\hat{\beta}_{6,2}$  is robust to bidirectional arcs strongly justified by prior knowledge but which have been omitted. In addition to directly affecting education and SES at age 30, parental SES is almost certainly confounded with these two offspring characteristics. At the very least there must be personality traits, independent of mental abilities, that influence attainment and are themselves genetically influenced (Figure 13). This would imply that we cannot trust either  $\hat{\beta}_{4,1}$  and  $\hat{\beta}_{5,1}$ ; these data by themselves do not allow us to say what the result of swapping households might have been on the attainments of this cohort. However, because the insertion of  $T_1 \leftrightarrow T_4$  and  $T_1 \leftrightarrow T_5$  does not create any new  $d$ -connecting paths between  $g$  and antiracism, these local breakdowns of identification do not affect our estimate of the  $T_2 \rightarrow T_6$  coefficient. After carrying out the  $d$ -separation tests

bearing on the assumption that there is no  $T_2 \leftrightarrow T_6$  arc, we can use regression to estimate the coefficient attached to the directed edge without bothering with the portions of the model that become unidentified when embedded in a more realistic supergraph.

Our conclusion is as follows. If we can somehow implement a manipulation to increase a child's level of  $g$  by age 11, it appears likely that the child will grow up to become a more racially tolerant adult. This extensive example has illustrated the distinctive features of the graphical approach to SEM, in particular highlighting how the testable implications of a causal model bear on specific substantive conclusions.

Some commentators have argued that any interesting high-level system will resist the analysis sketched above because of the corresponding DAG's *completeness* (Meehl & Waller, 2002; Freedman, 2004; Greenland, 2010); in such a graph, there are no  $d$ -separable pairs of nodes. This amounts roughly to the claim that, in any complex system, either everything affects everything or there are confounders that will never be identified. The antiracism example, however, suggests that the assumption of ubiquitous completeness may in fact be overly pessimistic. Perhaps further research will uncover more outcomes that are largely immune from the allegedly all-powerful contagion of SES.

Furthermore, in Part 4 I argue that there is at least one kind of causal system—the polygenic determination of a phenotype—where our prior knowledge is sufficient to dispel the intractability envisioned by skeptics of the graphical approach. *Quantitative or biometrical genetics* is the branch of population genetics concerned with the genetics of continuously varying traits (Lynch & Walsh, 1998; Bürger, 2000). Quantitative genetics has long been an integral part of personality research. It turns out that population genetics as a whole may be the basal theory needed to initiate the virtuous circle of “causal knowledge in, causal knowl-

edge out.” We now turn to the relevant aspects of this theory.

### 3.4 Concepts of Genetics

A genome-wide association study has an extremely simple design: a regression of the effect on the putative cause and a number of identically treated covariates. As we will shortly see, however, a replicable gene-trait *association* is nevertheless very strong evidence for gene-trait *causation*. As this degree of certainty is difficult to obtain in observational studies of comparable simplicity, gene hunting should be an attractive enterprise to personality researchers seeking a secure foothold for the traversal of the explanatory chain in Figure 11.

We now explore the insights that graphical reasoning brings to the search for specific DNA polymorphisms affecting ability variation.

#### 3.4.1 Foundations of Heritability

Many scientists believe that the evidence for the heritability of personality traits, provided by studies of twins and other kinships, is strong enough to justify studies aiming to find specific casual DNA variants (Bouchard & McGue, 2003). Here we elucidate the meaning of heritability from first principles, relying heavily on concepts that reappear in the later discussion of practical issues arising in gene-trait association studies.

#### 3.4.2 Ancestral Confounding

Suppose that we have a large number of loci in the genome associated with a trait of interest. Let  $p_i$  be the frequency of the allele to be counted at the  $i$ th such locus. Fisher 1999

expressed the *additive genetic variance* of the trait as

$$\text{Var}(A) = \sum_i 2p_i(1 - p_i)a_i\alpha_i, \quad (6)$$

where  $a_i$  and  $\alpha_i$  represent, respectively, the *average excess* and *average effect* of allelic substitution at the  $i$ th locus. The ratio of additive genetic variance to the total trait variance,

$$h^2 = \frac{\text{Var}(A)}{\text{Var}(Y)}, \quad (7)$$

is now known as the *heritability in the narrow sense*.

Fisher's manner of developing the concept of heritability, in particular his introduction of the variables that he called the "average excess" and "average effect," has struck some commentators as peculiar (Price, 1972; Falconer, 1985). It is my own belief, however, that Fisher's decision in *The Genetical Theory of Natural Selection* to base his discussion of heritability in terms of these variables was motivated in part by his recognition of the potential for gene-trait confounding. That is, the fact that different genotypes often correspond to different phenotypic values does not by itself show that the genotypic differences *cause* the phenotypic differences. It seems that this nicety was of great importance to Fisher. Therefore, in my recapitulation of the heritability concept, I emphasize how the distinction between confounding and causation enters into Fisher's two averages.

Geneticists refer to the confounding of genes and traits as *population structure* or *stratification*. A less formal term is the "chopstick gene syndrome": a gene showing an association with chopstick skill in a racially mixed sample is almost certainly not a gene "for" chopstick skill but rather a gene for black hair or yellow skin—or perhaps a gene where one allele

has drifted by chance to high frequency in East Asians. The apocryphal story of the geneticist misled by the chopstick gene illustrates how geographical subdivision can lead to gene-trait confounding. At some point in our evolutionary past, some humans split off from the rest of the African diaspora and became the ancestors of East Asians. Subsequently, natural selection and random genetic drift resulted in the divergence of allele frequencies among the branches of the diaspora. More recently, chopsticks were invented in China and diffused throughout what later became the Confucian belt. Thus, the ancestors of East Asians passed on both their genes and culture to their descendants, resulting in the confounding of genotypes and chopstick skill in mixed samples of East Asians and other peoples. Another consequence of geographical subdivision is substantial LD in the global human population; if a study participant has one allele that is at least somewhat associated with being East Asian, then it becomes more likely that the participant carries other such alleles.

Another mechanism of gene-trait confounding is assortative mating, the understanding of which is aided by a combination of genetic and graphical intuition. The following thought experiment closely follows a simulation study by Eaves (1979). Although the experiment does not accurately reflect how humans mate, it does reveal how a marital correlation arising from assortative mating falls under Pearl's addition to the correlational taxonomy. Suppose that upon reaching a given age, all members of a cohort form random opposite-sex pairings. If the man and woman within a random couple "hit it off," they go on to marry. Couples who are less fortunate break up, and the unmarried individuals may go through several more rounds of random pairing. Now suppose that after the first round we form a data matrix where each row corresponds to a randomly paired man and woman. The columns of this matrix record the trait values of each individual and also a binary variable indicating whether

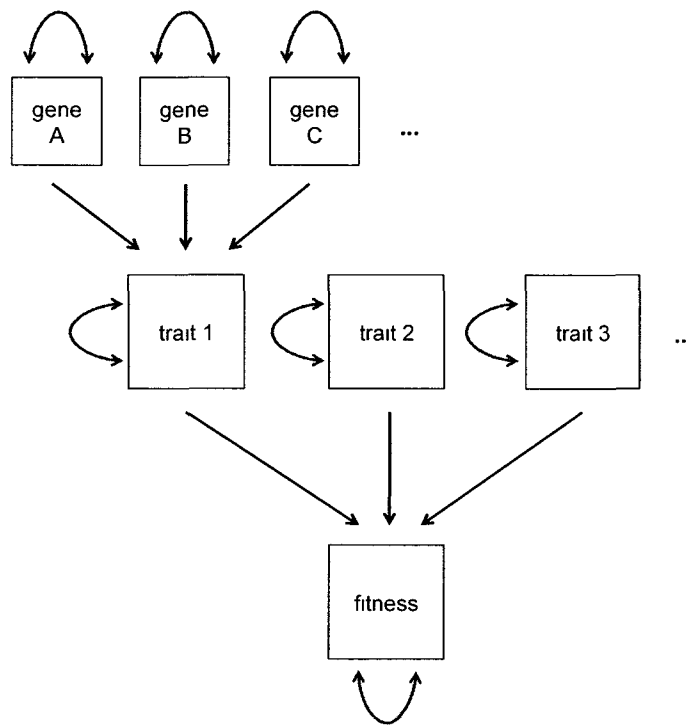


the two married at the end of the round. By stipulation, when considering all rows of this matrix, the correlation in trait value between male and female partners is not significantly different from zero. However, if we only consider those rows where the marriage indicator assumes the value one, any traits affecting the probability of marriage become correlated. That is, marriage is a collider.

This insight into the nature of assortative mating allows us to deduce that the trait-affecting genotypes of mother and father are  $d$ -connected because of conditioning on their common effect (mating). That is, those gametes with a trait-enhancing allele at one locus are more likely to be paired with gametes containing the enhancing alleles at other loci. Since the paternal and maternal contributions to a recombinant gamete will both tend to contain alleles with effects on the trait of the same sign, the coupling of same-sign alleles holds within gametes as well as between them (Crow & Kimura, 1970). Another way to describe the situation is to say that mating patterns among the ancestors of the population confound the genotypes at different loci affecting the same trait.

We now turn to the confounding property of natural selection. We can think of fitness (survival and reproduction) as a node with a multitude of directed edges converging on it from various phenotypes (Figure 17). Natural selection conditions on this node when deciding the ancestry (in the literal sense) of the offspring generation, and therefore all nodes ancestral (in the graphical sense) to fitness become  $d$ -connected. This theoretical finding implies that *all* functional sites in the genome are potentially in LD. In particular, if two loci affect a trait of which higher values are favored by selection, the enhancing allele at one locus will be associated with the depressing allele at the other.

Figure 17: A DAG representing the causal chain from genes to fitness. When considering selection bias in gene-trait association studies, we can simply relabel the bottom node as *appearance in the study*.



**Average Excess and Average Effect**      Can we isolate, either conceptually or experimentally, the causal effects of genetic differences at a single locus from the kinds of confounding mechanisms just described? Fisher's concepts of "average excess" and "average effect" appear to answer precisely this question. In his own words,

Let us now consider the manner in which any quantitative individual measurement, such as human stature, may depend upon the individual genetic constitution. We may imagine, in respect of any pair of alternative [alleles], the population divided into two portions, each comprising one homozygous type together with half of the heterozygotes, which must be divided equally between the two portions. The difference in average stature between these two groups may then be termed the average excess (in stature) associated with the gene substitution in question. This difference need not be wholly due to the single gene, by which the groups are distinguished, but possibly also to other genes statistically associated with it, and having similar or opposite effects. (Fisher, 1999, p. 30)

Fisher provided two definitions of the average effect. We first consider the definition that, although leading to great subtleties if pursued further, is more suggestive of the average effect's causal meaning:

[I]t is also necessary to give a statistical definition of a second quantity, which may be easily confused with that just defined, and may often have a nearly equal value, yet which must be distinguished from it in an accurate argument; namely the average effect produced in the population as genetically constituted, by the substitution of the one [allele] for the other. By whatever rules . . . the frequency

of different gene combinations, may be governed, the substitution of a small proportion of the [alleles] of one kind by the [alleles] of another will produce a definite proportional effect upon the average stature. The amount of the difference produced, on the average, in the total stature of the population, for each such gene substitution, may be termed the average effect of such substitution, in contra-distinction to the average excess as defined above. (Fisher, 1999, p. 31)

The basic notion is that a zygote is chosen at random from all those inheriting one allele (say  $\mathcal{A}_1$ ) from one parent (say the father).  $\mathcal{A}_1$  allele is then changed to  $\mathcal{A}_2$ , as if by mutation. The expected change in the phenotype  $Y$  at the time of measurement is then equal to the average effect. Thus, whereas all  $d$ -connecting paths between a genetic locus and the phenotype contribute to the average excess, a directed edge from gene to phenotype is necessary for a nonzero average effect at the focal locus. In Pearl's notation, then, the average excess is  $E[(Y | see(\mathcal{A}_2)) - (Y | see(\mathcal{A}_1))]$  whereas the average effect is  $E[Y | do(\mathcal{A}_2), see(\mathcal{A}_1)]$ . Any chopstick gene will show a positive average excess in the combined mixture of subpopulations but no average effect. All else being equal, under assortative mating the average excess will exceed the average effect; carriers of the two different alleles will tend to carry the alleles of like effect at other loci affecting the trait. Natural selection, on the other hand, will tend to reduce the average excess below the average effect.

The second definition of the average effect considers a multiple regression of the trait on all loci in the genome. The average effect of allelic substitution at the focal locus is equal to the partial regression coefficient of how many alleles, of the type to be counted (say  $\mathcal{A}_2$ ), are carried by the individual (Fisher, 1941). The two definitions of the average effect agree only in special circumstances (Falconer, 1985). Because Fisher does not even mention the

regression definition in the first edition of *The Genetical Theory*, it seems that he thought the causal definition to be more fundamental, and this is how we will treat it as well. The two definitions of the average effect coincide if gene action is purely additive. Both population-genetic theory and the available data suggest that for many traits pure additivity should be an acceptable approximation (Hill et al., 2008; Crow, 2010), and therefore the two definitions of the average effect should agree fairly well whenever a trait is determined by many loci of small effect.

Are the average excess and average effect ever equal? It can be shown that after many generations of random mating, in a broad sense that excludes not only assortative mating but natural selection and population structure, all LD and deviations from Hardy-Weinberg equilibrium will vanish (Crow & Kimura, 1970). Let us also assume that there are no confounders that affect the trait through environmental mediators. Then the focal locus is  $d$ -separated from all other causes of the trait, leaving a directed edge from the focal locus to the phenotype as the only means by which these two nodes are connected. That is, since the two population “portions, each comprising one homozygous type together with half of the heterozygotes,” do not differ in allele frequencies at any other loci, the difference in  $Y$  between them is attributable wholly to the average effect. The equivalence of the average excess and average effect under random mating is analogous to the equivalence of an observed difference and a causal effect under the randomization of treatment assignment. One naturally wonders about the degree to which Fisher’s thoughts on heritability and his work on experimental design stimulated each other.

In any event Equations (6) and (7) show that Fisher did indeed conceive of heritability as a causal concept.<sup>6</sup> Any genetic locus showing a positive average excess in the absence of an average effect contributes nothing to the heritability of the trait. Incidentally, the causal interpretation of heritability provides an intriguing perspective on Fisher’s Fundamental Theorem of Natural Selection. It is of course the central dogma of evolutionary biology that heritable variation in fitness leads to an increase in adaptation. The genius of Fisher’s theorem is that it captures this dogma in precise quantitative form: *the increase in the mean fitness of the population ascribable to the effect of natural selection on allele frequencies is equal to the additive genetic variance in fitness* Edwards (1994). Note how the distinction between association and causation affects the interpretation of Equation (6) as an expression for the change in the mean fitness. If a polymorphic locus is associated with fitness for any reason whatsoever—if its average excess is positive—one of its segregating alleles will increase in frequency. But this increase in allele frequency will only contribute to the adaptation of the species if substituting one allele for the other exerts a *causal* effect on fitness.

The discovery of the Fundamental Theorem was yet another blow struck by Fisher against his archenemy Pearson, who believed it was possible both to discard the notion of causality from science and to study evolution mathematically. If causality appears in the formulation of a phenomenon as fundamental as evolution by natural selection, then it surely cannot be a dispensable “fetish amidst the inscrutable arcana of modern science” (Pearson, 1911, p. xii).<sup>7</sup>

---

<sup>6</sup>The question arises as to why Fisher, a vigorous advocate of randomized experiments in agriculture and medicine, would commit himself to such strong causal claims regarding the heritability of height, fertility, economic productivity, and so forth on the basis of purely observational data. I myself do not believe that Fisher was being inconsistent here, but arguing this position is beyond the scope of this article.

<sup>7</sup>A caveat is in order. Under the regression definition of the average effect, the Fundamental Theorem of Natural Selection is a very general result that holds regardless of nonlinearities attributable to dominance,

### 3.4.3 Causal Inference in Gene-Trait Association Studies

The correlations between the trait values of relatives are functions of the narrow-sense heritability and other variance components, thus enabling the estimation of these parameters. The substantial heritabilities estimated for personality traits seem to justify attempts to map the specific DNA variants affecting them. The identification of these variants should lead to fundamental advances in our understanding of proximate mechanisms and the ultimate evolutionary forces shaping human personality (Figure 11). But recall the litany of potential confounding mechanisms that may result in a divergence of the average excess (which we can directly measure) from the average effect (which we want to know).

Given the number and complexity of potential confounding mechanisms, ruling out confounding at the level of individual genetic loci may seem to pose insurmountable difficulties. The litany of confounding mechanisms, however, is actually encouraging for the following reason. Since our knowledge of the mechanisms behind confounding is typically conjectural at best, in many cases we cannot say much about them. In contrast, the detail in which we can describe the population-genetic mechanisms behind confounding in gene-trait association studies reveals the depth of our knowledge in this domain. Exploiting our prior knowledge to characterize the relevant DAG, we can argue convincingly that all possible sources of confounding are controllable.

---

epistasis, gene-environment interaction, and so on. But before embracing the causal interpretation of the Fundamental Theorem, one would like to know how well the regression definition of the average effect isolates the causal effect of allelic substitution in a nonlinear setting. Falconer 1985 showed that for a single locus Fisher's "average causal effect" is proportional to the regression-based average effect, the constant factor depending on the extent of nonrandom mating, but I do not know whether such a simple and pleasing result holds in more complicated situations where nonlinearity and nonrandom mating cannot be described with just a few parameters. I suspect, however, that such a result does indeed hold. Finding such a result would be important because the Fundamental Theorem is intended to be an exact law of nature rather than a useful empirical approximation.

We can always rule out confounding as a source of gene-trait association in a within-family design (Laird & Lange, 2006). There exists a positive *within-family correlation* between variables  $X$  and  $Y$  if, across sibling pairs reared together, the sibling with the higher value of  $X$  also tends to have a higher value of  $Y$ . Personality researchers have long recognized that a within-family correlation presents stronger evidence for some causal relation than a correlation persisting after conditioning on familial background factors (Jensen & Sinha, 1993; Reiss et al., 1994; Turkheimer & Waldron, 2000; Beauchamp et al., 2011). We can now see that Pearl’s distinction between *seeing* and *doing* provides a rationale for this methodological principle. Whereas two unrelated individuals coincidentally sharing the same value of some familial variable may have come to that value for different reasons, siblings reared in the same home must have the same value for the same reasons. That is, within a family all background factors subsumed under “common” or “shared” environment have been *fixed* to some values, not merely *observed* to take on those values. It follows that any significant within-family correlation between  $X$  and  $Y$  cannot be the result of marginal or conditional confounding by factors that vary *across* families but not *within* them. In gene-trait association studies, a stronger claim is justified. When the putative causal variable is whether a sibling inherits  $\mathcal{A}_1$  or  $\mathcal{A}_2$  from a heterozygous parent, Mendel’s laws tell us that treatment assignment is literally at random. Since it is nature that performs this randomized experiment, we do not face the typical problem of deciding whether a human implementation of  $do(x)$  is really  $do(x, y, z)$ . *Given a reliable association between the within-family inheritance of a DNA marker and the phenotype, linkage between the marker and an authentic causal variant is the only viable explanation.* The recruitment of informative pedigrees can be quite difficult, however, and it is therefore desirable to seek other methods.



The fixing of genotype at fertilization does greatly restrict the class of alternative explanations for a gene-trait association. Obviously, we can rule out reverse causation; a manipulation of a person's phenotype will not induce mutation. More generally we can rule out any variable that follows fertilization in time. Given the complexity of the situation, however, this temporal restriction may initially fail to impress us. It is typically the absence of certain edges that enables effect identification, and in this case we have thousands of genetic loci that are each a sink for a dense network of evolutionary and historical forces. Oddly enough, it turns out that this case is also conducive to effect identification. Recall that Fisher's second definition of the average effect is the partial regression coefficient of allele count in the multiple regression of the trait on all loci in the genome. Recall also the theorem stated in Part 1: a partial regression coefficient gives an unbiased estimate of a causal effect in a linear system if the covariates  $d$ -separate all non-directed paths between putative cause and effect while including no descendant of the effect. Implicit in Fisher's definition, then, is a claim regarding the graphical properties of gene-trait confounders.

If the ancestral confounding consists of assortative mating or natural selection, then the average excess is contaminated by confounding because of LD between the focal locus and other loci. By including all other loci in the regression, we are intercepting each and every non-directed path to the phenotype through these non-focal loci, thereby justifying the *statistically* defined average effect as a *causal* effect. However, if the ancestral confounding arises from geographical subdivision or some other form of population structure, there may be non-directed paths mediated by environmental variables that have not been measured. A rather special feature of population structure allows us to overcome this difficulty: the *entire* genome is subject to the selective and stochastic divergence of allele frequencies among

subpopulations after the splintering of their ancestral population. Thus, as the number of loci entering the regression becomes very large, they become a perfect proxy for the subpopulation (or location in continuous ancestral space) to which a study participant belongs. By partialing out all loci in the genome, then, we are in effect partialing out the ancestral events confounding the gene and the trait.

From our discussion of population structure, assortative mating, and natural selection, we generalize as follows: *Every possible confounder of gene and trait has the property of being mediated by another genetic locus or sending directed paths to thousands of genetic loci. This property allows us to control gene-trait confounding by conditioning on all other loci in the genome.* These statements are not at all rigorous, and examples could be contrived to defeat them. Nevertheless the examples of gene-trait confounding that we have examined suggest that the principle is quite robust. When combined judiciously with within-family designs, studies of nominally unrelated individuals controlling for genome-wide background should be a reliable tool for pinpointing the causal effects of genetic differences.

Since the number of independently segregating regions of the genome will usually exceed the sample size, some kind of proxy for the entire genome is typically employed. Successful tools for this purpose have so far included the principal components of the genotype matrix (Price et al., 2006; McVean, 2009) and computational simplifications of treating all genotyped markers as a random effect (Kang et al., 2010). As sample sizes continue to increase and the transition to whole-genome sequencing accelerates, the ideal of actually conditioning on all loci in the genome will be ever more closely approached.

It is remarkable that observational research employing so simple a design—regression of the effect on the putative cause and a number of undifferentiated covariates—can produce

such trustworthy causal inferences *in principle*. The addendum is necessary because of the problems introduced by *selection bias*, which occurs whenever a trait being studied is itself a cause of appearance in the study. Since an individual genetic variant is likely to have a very small effect, extremely large samples are required to detect it (Manolio et al., 2009). Gene hunters may have to sacrifice methodological perfectionism to attain the necessary scale. “Personal genomics” studies, drawing upon large all-volunteer samples, have reported associations of genetic variants with hair morphology, freckling, asparagus anosmia, photic sneeze reflex, and Parkinson’s disease (Eriksson et al., 2010; Do et al., 2011). This approach will soon be extended to encompass whole-genome sequencing of all-volunteer samples exceeding 100,000 in size (Lunshof et al., 2010), and the not-too-distant future may bring even greater orders of magnitude. We can see from Figure 17, however, that the effect of selection bias on the divergence between the average excess and average effect is qualitatively the same as that of natural selection.<sup>8</sup> If we imagine that all individuals not volunteering for a given study subsequently perish or fail to reproduce, then the analogy to natural selection is exact. The quantitative effect of selection bias on LD will typically be much stronger than that of natural selection for several reasons: (1) personality traits such as intelligence, openness, and religiosity will have much stronger effects on study participation than on fitness itself; (2) recombination has no opportunity to reduce this source of LD; and (3) any environmental effect on the trait will be negatively correlated with the number of enhanc-

---

<sup>8</sup>This device of treating appearance in a study as a node with edges connecting it to the variables being studied can be greatly generalized to address all problems of missing data (Meredith, 1993; Schafer & Graham, 2002; Little & Rubin, 2002). Some researchers may find the judgment of whether one variable causes another to be more natural than consideration of the conditional probabilities arising in the potential-outcome framework; in any case the two approaches are mathematically equivalent. See Daniel, Kenward, Cousens, and De Stavola (2011) and Barenboim and Pearl (2011) for discussion.

ing alleles at a trait-affecting locus. Consequently, there is reduced power to detect loci with true effects, an underestimation of the average effect at any detected locus, and a surfeit of false-positive loci affecting other traits that are causes of study participation.

Because of the *d*-connections between trait-affecting loci and environmental disturbances, all other loci in the genome do not constitute an adequate *d*-separating set in the presence of selection bias. It might seem from Figure 17 that we can control selection bias by including all relevant *traits* as covariates. Unlike the genes, however, the traits may be causally ordered. If there are colliders and mediators among the traits in the covariate set, then conditioning on these traits invites the problems detailed at length in Part 1. In fact, the lack of a causal order among different loci in the genome is what makes the genomic background such an effective shield against confounding, and we might fairly say that it is this graphical property that gives gene-trait association studies of unrelated individuals their special character with respect to the warrant of causal inferences.

Nevertheless the measurement of those traits likely to affect appearance in a gene-trait association study appears to be a desirable methodological safeguard. Since selection bias may distort the factorial structure of personality measurements (Meredith, 1993), extra care must be taken to ensure their reliability. If a DNA marker shows an association with these traits, investigators will at least be alerted to the possibility that an additional association with some focal trait may be the result of an unblocked collision at study participation. As mentioned, personality traits are certain to be among the most important causes of volunteering. If the association with the focal trait is the only one remaining after conditioning on the traits likely to affect study appearance, the investigators may tentatively hypothesize that the association reflects a genuine causal effect on the focal trait. Any firmer conclusion must

await replication, at perhaps a less stringent significance threshold, in a study where personal characteristics have a negligible impact on participation.

### 3.5 Conclusion

It is a testament to Fisher's intuition that he was able to do so much with the concepts of association and causation at a time when the distinction between was poorly understood and in fact scorned by the leading intellects of the day. Indeed, his Fundamental Theorem of Natural Selection was the first (and, so far, the only known) law of nature explicitly depending on a distinction between association and causation. Wright's diagrammatic approach to cause and effect serves as a convenient conceptual bridge toward Pearl's graphical formalization of causality, which has greatly extended the innovations in causal reasoning developed by both of the population-genetic pioneers.

The fruitfulness of Pearl's graphical theory when applied to the problems discussed in this article bear out its utility to personality psychology. Perhaps the most surprising instance of the theory's fruitfulness concerns the role of colliders. Although obscure before Pearl's seminal work, this role turns out to be obvious in retrospect and a great aid to the understanding of many seemingly unrelated problems. This article has surely only scratched the surface of the ramifications following from our recognition of colliders.

The absence of a formal vehicle for causal notions, however, cannot be a full explanation for why the debates over the causes of personality have been so fractious. We have employed the trait of general intelligence ( $g$ ) in many of our examples, and as a result we have seen that there must be few important aspect of human affairs falling outside its surmised influence. Status attainment, health, mortality, mating preferences, even beliefs regarding

how society should be organized—all plausibly affected by *g* (and perhaps other personality traits as well). It is little wonder then that the questions regarding the causes of individual differences have attracted so much controversy. Some have supposed that if these causes can be traced to environmental sources, we should easily be able to manipulate human personality through appropriate educational or social interventions. In contrast, “hereditarianism” has been condemned as “damaging . . . and . . . malicious . . . for it shatters hope that science can improve the human condition” (Glymour, 1997, p. 278). Such thinking, however, may be premature. Given continuing advances in reproductive technology, it is not at all clear that a strong genetic influence on intelligence is incompatible with its manipulability. The implication for researchers is that we should formulate and test causal theories without prejudging their consequences for how we might distribute the gift of rationality more equitably than chance and nature have seen fit to do.

To the extent that manipulating the genetic causes of personality is discussed at all, the possibility is typically seen as being ethically problematic (Goldstein, 2011a). This reveals a rather curious asymmetry, since manipulating *environmental* causes of personality rarely provokes any concern. By pointing out this one-sidedness, I do not mean to imply that the deliberate molding of personality raises no ethical issues. But recognizing a democracy of causes must unavoidably change the entire tone of the coming discussion.

These extra-scientific concerns remind us that we value causal knowledge for its leverage in manipulating the world to suit our own purposes. Thus, with respect to manipulability, whether “hereditarianism” is true is in fact a secondary issue. More important is whether causality is true. And this is the present article’s central message: causality is true.

## General Conclusion

I now briefly discuss what has been learned in the three papers regarding the issues raised in the General Introduction and the next steps in the overall research program.

The first paper described a genome-wide association studies of over 100 physical and behavioral phenotypes failing to find any loci associated with the major personality traits. This result is consistent with the genetic architecture of a typical personality trait consisting of many loci of small effect—a conclusion also supported by several reports that have been recently published or are still in press (Davis et al., 2010; de Moor et al., 2011; Davies et al., 2011). I offer the hypothesis that since stabilizing selection tends to eliminate variants of large effect, we will tend to observe such variants contributing substantially to the variability of a quantitative trait only in unusual circumstances (e.g., a selection pressure that is large relative to the initial genetic variation). I state this hypothesis rather informally, and it may be worthwhile to flesh it out in explicit mathematical form in order to clarify the meanings of terms like “large.”

My argument that the typical genetic architecture of a quantitative trait consists of many “infinitesimal” loci commits to one side of an ongoing debate regarding the sources of the “missing heritability” that genome-wide association studies have not yet accounted for. The other side, inspired by laboratory studies of model organisms and theoretical arguments regarding “synthetic” associations, emphasizes the possibility that a given trait may be affected by relatively few loci of large effect. In the Discussion of the first paper, I summarize the reasons why the available empirical evidence is more consistent with the infinitesimal hypothesis than with the “large effect” hypotheses. However, a position in this debate has

consequences for how to best advance this field only in the short term. With respect to the long term, I agree with the closing sentence of Goldstein's (2011b) reply to the critiques of his arguments for synthetic associations: "Sequencing is and should be the future of discovery genetics" (p. 3). A combination of sequencing, cross-ethnic comparisons, and functional studies should do much to resolve the current uncertainties regarding the genetic architectures of quantitative traits.

Fully exploiting the available and developing technology for resolving the genetic architecture of a given personality trait will require sample sizes much larger than those typically employed in behavioral research. I am currently involved in a number of discussions and active collaborations toward this end and expect that research in this area will only continue to grow.

The second paper presented evidence supporting the partition of RT into several processing stages, only one of which is correlated with IQ. This central "response selection" stage processes its inputs in serial fashion and contains a stochastic accumulation of evidence toward one of a few discrete alternatives. One potential criticism of our RT studies is that the unreplicated findings in our small and unrepresentative sample do not adequately support our strong claims regarding our detailed mechanistic model (Tversky & Kahneman, 1971; Ioannidis, 2005). This is a just criticism, and thus I emphasize the need to replicate the key findings in larger and more representative samples. Furthermore, we should examine the effects of prolonged practice and attempt to rule out reverse causation and trivial confounding as alternative explanations, employing the Pearlian logic described in the third paper to inform specific study designs. After replication of the basic findings and causal validation, we can then begin tying the mechanisms posited in the unified RT model to both underly-



ing neural properties and overlying complex cognitive processes. By having identified very distinctive features of the IQ-associated processing stage (namely its stochasticity, seriality, and central temporal position), we have greatly enabled these integrative efforts. These features place strong constraints on higher-level theories and give clear indications of what to look for when delving into the brain. Also, one way to alleviate the burden of sample size in genome-wide association studies of personality traits is to measure endophenotypes lying causally closer to the genes, and the validation of central processing in RT as a causal mediating variable may do much to advance this goal for genetic studies of intelligence.

The aspects of Pearl's graphical theory of causality presented in the third paper provide an essential foundation for the entire endeavor of attempting to discover the causes and consequences of personality. In this summary of what was learned in the paper, I will keep my description at a relatively abstract level. Pearl's theory accomplishes what many Edwardian scientists, including Bertrand Russell and Karl Pearson, thought was impossible: it captures human intuitions about causality in the form of mathematical axioms. These axioms allow us to state in a formal language that breaking the glass of water *will* make the floor wet—but that making the floor wet *will not* break the glass of water. Remarkably there was no way to express this distinction in the traditional language of probability theory. Now it is clear that we do not need a formal language to know that wet floors do not break glasses! But once we are equipped with a formal language that can express this notion, we can use it in more complicated situations to derive highly nontrivial conclusions that are beyond the reach of the unaided intellect. Such derivations were demonstrated in the examples and data reanalysis presented in the third paper.

It should be understood that the graphical theory does not magically turn a matrix of cor-

relations (or, more generally, a joint probability distribution) into a list of quantitative causal effects. Causal inferences always take a conditional form: *if* we know or assume certain things to be true, *then* certain casual conclusions inevitably follow. The relevant background assumptions can include parsimony, temporal order, a real-world experimental manipulation conforming to a mathematically idealized manipulation, and so on. One reason why I discuss genome-wide association studies at some length in the third paper is that our substantial prior knowledge of genetics allows us to justify some very powerful assumptions; it is these assumptions that counter Turkheimer's (2008) attempt to analogize genome-wide association studies to various failures of observational methodology in the social sciences. For example, when discussing within-family designs, Turkheimer does not point out that according to Mendel's Law of Segregation, the allele that a heterozygous parent passes on to an offspring is determined *randomly*. Thus, within-family designs in genome-wide association studies do not merely fix (a subset of) potential confounders; the level of the putative causal variable is literally randomized, warranting a very high grade of confidence in causal inferences. Although this benefit of natural randomization is not available in genome-wide association studies of unrelated individuals, I go on to discuss reasons why observed associations in such studies are still likely to reflect genuine causal effects.

Overall, then, what lessons emerge from the three papers collected in this dissertation? Although much work lies ahead, I believe that the papers demonstrate, by both theoretical arguments and empirical proof of principle, the feasibility of a research program intent on substantial advances in the mechanistic understanding of personality variation. Thanks to technological advances in measurement resolution, closer contact with cognate branches of psychology, and a formal vehicle for causal ideas, I believe that seeking the causes of

individual differences in genetics, neural properties, and information-processing mechanisms has become a matter of ordinary scientific ingenuity rather than an endeavor suffering from inevitable epistemological defects.

## References

- Aitken Harris, J., Vernon, P. A., & Jang, K. L. (2005). Testing the differentiation of personality by intelligence hypothesis. *Personality and Individual Differences*, *38*, 277–286.
- Anderson, C. A., Soranzo, N., Zeggini, E., & Barrett, J. C. (2011). Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biology*, *9*, e1000580.
- Anderson, J. C. & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.
- Anthony, D. W. (2007). *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World*. Princeton, NJ: Princeton University Press.
- Ashton, M. C. (2007). *Individual Differences and Personality*. Burlington, MA: Elsevier.
- Ashton, M. C. & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality*, *15*, 327–353.
- Ashton, M. C. & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality*, *19*, 5–24.
- Baars, B. J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. New York, NY: Oxford University Press.
- Baddeley, A. (2007). *Working Memory, Thought, and Action*. New York, NY: Oxford University Press.
- Barenboim, E. & Pearl, J. (2011). Controlling selection bias in causal inference.
- Bartholomew, D. J. (2004). *Measuring Intelligence: Facts and Fallacies*. Cambridge, UK: Cambridge University Press.
- Batty, G. D., Wennerstad, K. M., Davey Smith, G., Gunnell, D., Deary, I. J., Tynelius, P., & Rasumussen, F. (2009). IQ in early adulthood and mortality by middle age. *Epidemiology*, *20*, 100–109.
- Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., C., L., Li, J. C., Liang, Y., McCormack, M., et al. (2010). Natural selection on *EPAS1* (*HIF2 $\alpha$* ) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences USA*, *107*, 11459–11464.

- Beauchamp, J. P., Cesarini, D., Johannesson, M., Lindqvist, E., & Apicella, C. (2011). On the sources of the height-intelligence correlation: New insights from a bivariate ACE model with assortative mating. *Behavior Genetics, 41*, 242–252.
- Bellwood, P. S. (2005). *First Farmers: Origins of Agricultural Societies*. Malden, MA: Blackwell.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–218.
- Bouchard, T. J. & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavior Genetics, 31*, 243–273.
- Bouchard, T. J. & McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology, 54*, 4–45.
- Box, J. F. (1978). *R.A. Fisher: The Life of a Scientist*. New York, NY: Wiley.
- Brito, C. & Pearl, J. (2002). A new identification condition for recursive models with correlated errors. *Structural Equation Modeling, 9*, 459–474.
- Bulmer, M. G. (1971). The effect of selection on genetic variability. *American Naturalist, 105*, 201–211.
- Bürger, R. (2000). *The Mathematical Theory of Selection, Recombination, and Mutation*. Chichester, UK: Wiley.
- Burt, C. (1940). *The Factors of the Mind*. London, UK: University of London Press.
- Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G., & Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics, 37*, 868–872.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431.
- Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.
- Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Experimental game theory and behavior genetics. *Annals of the New York Academy of Sciences, 1167*, 66–75.

- Chabris, C. F. (2007). Cognitive and neurobiological mechanisms of the Law of General Intelligence. In M. J. Roberts (Ed.), *Integrating the Mind* (pp. 449–491). Hove, UK: Psychology Press.
- Chabris, C. F., Laibson, D. I., & Schuldt, J. P. (2008). Intertemporal choice. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed.). (pp. 536–542). London, UK: Palgrave Macmillan.
- Chabris, C. F. & Simons, D. (2010). *The Invisible Gorilla: And Others Ways Our Intuitions Deceive Us*. New York, NY: Crown.
- Cochran, G. & Harpending, H. (2009). *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York, NY: Basic Books.
- Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N. (2007). *Variation in Working Memory*. New York, NY: Oxford University Press.
- Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological Science*, *19*, 1110–1117.
- Costa, P. T. & McCrae, R. R. (1992). *NEO Personality Inventory–Revised (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.
- Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B*, *365*(1544), 1241–1244.
- Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York, NY: Harper and Row.
- Daniel, R. M., Kenward, M. G., Cousens, S. N., & De Stavola, B. L. (2011). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*.
- Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., Hellard, S. L., Christoforou, A., Luciano, M., et al. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*.

- Davis, O. S. P., Butcher, L. M., Docherty, S. J., Meaburn, E. L., Curtis, C. J. C., Simpson, M. A., Schalkwyk, L. C., & Plomin, R. (2010). A three-stage genome-wide association study of general cognitive ability: Hunting the small effects. *Behavior Genetics, 40*, 31–45.
- De Jong, R. (1993). Multiple bottlenecks in overlapping task performance. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 965–980.
- de Moor, M. H. M., Costa, P. T., Terracciano, A., Krueger, R. F., de Geus, E. J. C., Toshiko, T., Penninx, B. W. J. H., Esko, T., Madden, P. A. F., & Derringer, J. (2011). Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry, 16*, 1023–1034.
- Deary, I. J. (2001). Human intelligence differences: Towards a combined experimental-differential approach. *Trends in Cognitive Sciences, 5*, 164–170.
- Deary, I. J., Batty, G. D., & Gale, C. R. (2008). Bright children become enlightened adults. *Psychological Science, 19*, 1–6.
- Deary, I. J., Der, G., & Ford, G. (2001). Reaction times and intelligence differences: A population-based cohort study. *Intelligence, 29*, 389–399.
- Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience, 11*, 201–211.
- Dehaene, S. (1996). The organization of brain activations in number comparison: Event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience, 8*, 47–68.
- Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In *Attention and Performance XXII: Sensorimotor Foundations of Higher Cognition* (pp. 527–574). New York, NY: Oxford University Press.
- Dehaene, S. (2008). Conscious and nonconscious processes: Distinct forms of evidence accumulation? In C. Engel & W. Singer (Eds.), *Better Than Conscious? Decision Making, the Human Mind, and Implications for Institutions* (pp. 21–49). Cambridge, MA: MIT Press.
- Dell'Acqua, R., Jolicoeur, P., Vespignani, F., & Toffanin, P. (2005). Central processing overlap modulates P3 latency. *Experimental Brain Research, 165*, 54–68.

- Dennett, D. C. (1991). *Consciousness Explained*. New York, NY: Little and Brown.
- Dickens, W. T. & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 356–369.
- Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., & Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*, *8*, e1000294.
- Do, C. B., Tung, J. Y., Dorfman, E., Kiefer, A. K., Drabant, E. M., Francke, U., Mountain, J. L., Goldman, S. M., Tanner, C. M., Langston, J. W., et al. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genetics*, *7*, e1002141.
- Dolgin, E. (2010). Personalized investigation. *Nature Medicine*, *16*, 953–955.
- Duchaine, B. & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*, 576–585.
- Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin and Review*, *16*, 1026–1036.
- Eaves, L. J. (1979). The use of twins in the analysis of assortative mating. *Heredity*, *43*, 399–409.
- Edwards, A. W. F. (1994). The Fundamental Theorem of Natural Selection. *Biological Reviews*, *69*, 443–474.
- Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., Goldman, D., & Weinberger, D. R. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences USA*, *98*, 6917–6922.
- Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., & Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics*, *6*, e1000993.
- Eyre-Walker, A. (2010). Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences USA*, *107*, 1752–1756.



- Falconer, D. S. (1985). A note on Fisher's 'average effect' and 'average excess'. *Genetical Research*, 46, 337–347.
- Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics* (4th ed.). Harlow, UK: Pearson.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications Vol. I* (3rd ed.). New York, NY: Wiley.
- Ferreira, V. S. & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 1187–1199.
- Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52, 399–433.
- Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford, UK: Clarendon.
- Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, 11, 53–63.
- Fisher, R. A. (1966). *The Design of Experiments* (8th ed.). New York, NY: Hafner.
- Fisher, R. A. (1970). *Statistical Methods for Research Workers* (14th ed.). New York, NY: Hafner.
- Fisher, R. A. (1999). *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford, UK: Oxford University Press.
- Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
- Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review*, 28, 267–293.
- Frey, M. C. & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the SAT and general cognitive ability. *Psychological Science*, 15, 373–378.
- Friedman, H. S., Tucker, J. S., Tomlinson-Keasey, C., Schwartz, J. E., Wingard, D. L., & Criqui, M. H. (1993). Does childhood personality predict longevity? *Journal of personality and Social Psychology*, 65, 176–185.

- Frost, P. (2006). European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior*, 27, 85–103.
- Gallacher, J., Bayer, A., Dunstan, F., Yarnell, J., Elwood, P., & Ben-Shlomo, Y. (2009). Can we understand why cognitive function predicts mortality? Results from the Caerphilly Prospective Study (CaPS). *Intelligence*, 37, 535–544.
- Gallistel, C. R. & Gelman, R. (2005). Mathematical cognition. In K. J. Holyoak & R. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 559–588). Cambridge, UK: Cambridge University Press.
- Gillespie, N. A. & Martin, N. G. (2005). Direction of causation models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 496–499). Chichester, UK: Wiley.
- Glymour, C. (1997). Social statistics and genuine inquiry: Reflections on *The Bell Curve*. In *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve* (pp. 257–280). New York, NY: Springer.
- Gold, J. I. & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, 404, 390–394.
- Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574.
- Goldstein, D. B. (2011a). Growth of genome screening needs debate. *Nature*, 476, 27–28.
- Goldstein, D. B. (2011b). The importance of synthetic associations will only be resolved empirically. *PLoS Biology*, 9, e1001008.
- Gosso, M. F., van Belzen, M., de Geus, E. J. C., Polderman, J. C., Heutink, P., Boomsma, D. I., & Posthuma, D. (2006). Association between the *CHRM2* gene and intelligence in a sample of 304 Dutch families. *Genes, Brain and Behavior*, 5, 577–584.
- Gottfredson, L. S. & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science*, 13, 1–4.
- Gould, S. J. (1981). *The Mismeasure of Man*. New York, NY: Norton.
- Grasman, R. P. P. P., Wagenmakers, E. J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology*, 53, 55–68.

- Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316–322.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.
- Greenland, S. (2010). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (pp. 365–382). London, UK: College Publications.
- Gribbin, J. (2002). *Science: A History*. London, UK: Penguin.
- Guttman, L. & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence*, *15*, 79–103.
- Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S. E., Hu, F. B., Duffy, D. L., Zhao, Z. Z., et al. (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genetics*, *4*, e1000074.
- Hanes, D. P. & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*, *274*, 427–430.
- Hastings, A. (1990). Second-order approximations for selection coefficients at polygenic loci. *Journal of Mathematical Biology*, *28*, 475–483.
- Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, *9*, 467–479.
- Hesselmann, G., Flandin, G., & Dehaene, S. (2011). Probing the cortical network underlying the psychological refractory period: A combined EEG-fMRI study. *NeuroImage*, *56*, 7585–7598.
- Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, *4*, e1000008.
- Hofstadter, D. R. & the Fluid Analogies Research Group (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York, NY: Basic Books.
- Holland, P. (1995). Some reflections on Freedman's critiques. *Foundations of Science*, *1*, 50–57.

- Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., Magnusson, O. T., Helgason, A., Saemundsdottir, J., Gylfason, A., et al. (2011). A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nature Genetics*, *43*, 316–320.
- Hommel, B. (1998). Automatic stimulus-response translation in dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1368–1384.
- Humphreys, L. G. (1994). Intelligence from the standpoint of a (pragmatic) behaviorist. *Psychological Inquiry*, *5*, 179–192.
- Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review*, *85*, 109–130.
- Hunt, E. (2005). Information processing and intelligence: Where we are and where we are going. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and Intelligence: Identifying the Mechanisms of Mind* (pp. 1–25). Cambridge, UK: Cambridge University Press.
- Hunt, E., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology*, *7*, 194–227.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.
- Jablonski, N. G. & Chaplin, G. (2010). Human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences USA*, *107*, 8962–8968.
- Jackson, D. N. (1998). *Multidimensional Aptitude Battery II* (2nd ed.). Port Huron, MI: Sigma Assessment Systems.
- Jensen, A. R. (1987a). Individual differences in the Hick paradigm. In P. A. Vernon (Ed.), *Speed of Information Processing and Intelligence* (pp. 101–175). Norwood, NJ: Ablex.
- Jensen, A. R. (1987b). Process differences and individual differences in some cognitive tasks. *Intelligence*, *11*, 107–136.
- Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.
- Jensen, A. R. (2006). *Clocking the Mind: Mental Chronometry and Individual Differences*. Oxford, UK: Elsevier.

- Jensen, A. R., Cohn, S. J., & Cohn, C. M. G. (1989). Speed of information processing in academically gifted youths and their siblings. *Personality and Individual Differences, 10*, 29–33.
- Jensen, A. R. & Sinha, S. N. (1993). Physical correlates of human intelligence. In P. A. Vernon (Ed.), *Biological Approaches to the Study of Human Intelligence* (pp. 139–242). Norwood, NJ: Ablex.
- Jiang, Y., Saxe, R., & Kanwisher, N. (2004). Functional magnetic resonance imaging provides new constraints on theories of the psychological refractory period. *Psychological Science, 15*, 390–396.
- Johnson-Laird, P. N. (2006). *How We Reason*. New York, NY: Oxford University Press.
- Jung, R. E. & Haier, R. J. (2007). The parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences, 30*, 135–154.
- Kamienkowski, J. E., Pashler, H., Dehaene, S., & Sigman, M. (2011). Effects of practice on task architecture: Combined evidence from interference experiments and random-walk models of decision making. *Cognition, 119*, 81–95.
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics, 42*, 348–354.
- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press.
- Knafo, A., Israel, S., Darvasi, A., Bachner-Melman, R., Uzefovsky, F., Cohen, L., Feldman, E., Lerer, E., Laiba, E., Raz, Y., et al. (2008). Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes, Brain and Behavior, 7*, 266–275.
- Koenig, L. B., McGue, M., Krueger, R. F., & Bouchard, T. J. (2005). Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality, 73*, 471–488.
- Laird, N. M. & Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics, 7*, 385–394.

- Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution*, 33, 402–416.
- Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. W., Fernando, R., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467, 832–838.
- Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. Chichester, UK: Wiley.
- Lee, T. S., Yang, C. F., Romero, R. D., & Mumford, D. (2002). Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience*, 5, 589–597.
- Li, W., Piech, V., & Gilbert, C. D. (2008). Learning to link visual contours. *Neuron*, 57, 442–451.
- Lien, M.-C., Ruthruff, E., & Johnston, J. C. (2006). Attentional limitations in doing two tasks at once. *Current Directions in Psychological Science*, 15, 89–93.
- Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.
- Liu, F., Wollstein, A., Hysi, P. G., Ankra-Badu, G. A., Spector, T. D., Park, D., Zhu, G., Larsson, M., Duffy, D. L., Montgomery, G. W., et al. (2010). Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genetics*, 6, e1000934.
- Logan, G. D. & Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychological Review*, 108, 393–434.
- Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Lubinski, D. & Dawis, R. V. (1995). *Assessing Individual Differences in Human Behavior*. Palo Alto, CA: Consulting Psychologists.
- Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

- Lunshof, J. E., Bobe, J., Aach, J., Angrist, M., Thakuria, J. V., Vorhaus, D. B., Hoehe, M. R., & Church, G. M. (2010). Personal genomes in progress: From the Human Genome Project to the Personal Genome Project. *Dialogues in Clinical Neuroscience, 12*, 47–60.
- Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences USA, 104*, 8597–8604.
- Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer.
- MacArthur, D. G., Seto, J. T., Raftery, J. M., Quinlan, K. G., Huttley, G. A., Hook, J. W., Lemckert, F. A., Kee, A. J., Edwards, M. R., Berman, Y., et al. (2007). Loss of *ACTN3* gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nature Genetics, 39*, 1261–1265.
- MacCorquodale, K. & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55*, 95–107.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorf, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature, 461*, 747–753.
- Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology, 64*, 17–24.
- Marois, R. & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences, 9*, 296–305.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.
- Maynard Smith, J. & Price, G. R. (1973). The logic of animal conflict. *Nature, 246*, 15–18.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., Hirschhorn, J. N., et al. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics, 9*, 356–369.
- McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.
- McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Hillsdale, NJ: Erlbaum.

- McDonald, R. P. (1996). Consensus emergens: A matter of interpretation. *Multivariate Behavioral Research*, 31, 663–672.
- McDonald, R. P. (2002). What can we learn from the path equations?: Identifiability, constraints, equivalence. *Psychometrika*, 67, 225–249.
- McDonald, R. P. (2003). Behavior domains in theory and practice. *Alberta Journal of Educational Research*, 49, 212–230.
- McDonald, R. P. (2004). The specific analysis of structural equation models. *Multivariate Behavioral Research*, 39, 687–713.
- McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, 5, 675–686.
- McDonald, R. P. & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82.
- McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5, e1000686.
- Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota Studies in the Philosophy of Science Vol. IV* (pp. 373–402). Minneapolis, MN: University of Minnesota Press.
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1, 108–141.
- Meehl, P. E. & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods*, 7, 283–300.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.
- Meyer, D. E. & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: I. Basic mechanisms. *Psychological Review*, 104, 3–65.
- Miller, L. T. & Vernon, P. A. (1992). The general factor in short-term memory, intelligence, and reaction time. *Intelligence*, 16, 5–29.



- Moyer, R. S. & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520.
- Mulaik, S. A. (2005). Looking back on the indeterminacy controversies in factor analysis. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics: A Festschrift for Roderick P. McDonald* (pp. 174–206). Mahwah, NJ: Erlbaum.
- Mulaik, S. A. (2010). *Foundations of Factor Analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Mulaik, S. A. & McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, *43*, 177–192.
- Murray, C. (2002). IQ and income inequality in a sample of sibling pairs from advantaged family backgrounds. *American Economic Review*, *92*, 339–343.
- Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., et al. (2010). From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature*, *466*, 714–719.
- Muthén, B. O. (1987). Response to Freedman’s critique of path analysis: Improve credibility by better methodological training. *Journal of Educational and Behavioral Statistics*, *12*, 178.
- Need, A. C., Attix, D. K., McEvoy, J. M., Cirulli, E. T., Linney, K. N., Wagoner, A. P., Gumbs, C. E., Giegling, I., Möller, H. J., Francks, C., et al. (2008). Failure to replicate effect of *Kibra* on human memory in two large cohorts of European origin. *American Journal of Medical Genetics B*, *147*, 667–668.
- Nisbett, R. E. (2009). *Intelligence and How to Get It: Why Schools and Cultures Count*. New York, NY: Norton.
- Olson, J. M., Vernon, P. A., Harris, J. A., & Jang, K. L. (2001). The heritability of attitudes: A study of twins. *Journal of Personality and Social Psychology*, *80*, 845–860.
- O’Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, *314*, 91–94.
- Papassotiropoulos, A., Stephan, D. A., Huentelman, M. J., Hoerdli, F. J., Craig, D. W., Pearson, J. V., Huynh, K. D., Brunner, F., Corneveaux, J., Osborne, D., et al. (2006). Common *Kibra* alleles are associated with human memory performance. *Science*, *314*, 475–478.

- Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 358–377.
- Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220–244.
- Pashler, H. (1998). *The Psychology of Attention*. Cambridge, MA: MIT Press.
- Pashler, H. & Johnston, J. C. (1989). Chronometric evidence for central postponement in temporally overlapping tasks. *Quarterly Journal of Experimental Psychology*, *41A*, 19–45.
- Pashler, H. & Johnston, J. C. (1998). Attentional limitations in dual-task performance. In H. Pashler (Ed.), *Attention* (pp. 155–189). Hove, UK: Psychology Press.
- Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*, 768–774.
- Payton, A., Holland, F., Diggle, P., Rabbitt, P., Horan, M., Davidson, Y., Gibbons, L., Worthington, J., Ollier, W., & Pendleton, N. (2003). Cathepsin D exon 2 polymorphism associated with general intelligence in a healthy older population. *Molecular Psychiatry*, *8*, 14–18.
- Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, *27*, 226–284.
- Pearl, J. (2004). Robustness of causal claims. In M. Chickering & J. Halpern (Eds.), *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence* (pp. 446–453). Arlington, VA: AUAI Press.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). New York, NY: Cambridge University Press.
- Pearl, J. & Verma, T. (1987). The logic of representing dependencies by directed acyclic graphs. In *Proceedings of the Sixth National Conference on AI* (pp. 374–379). Seattle, WA: AAAI Press.
- Pearson, K. (1911). *The Grammar of Science* (3rd ed.). London, UK: Black.
- Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality*, *21*, 549–587.

- Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, *19*, 826–837.
- Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral Genetics* (5th ed.). New York, NY: Worth Publishers.
- Plomin, R., Turic, T. M., Hill, L., Turic, D. E., Stephens, M., Williams, J., Owen, M. J., & O'Donovan, M. C. (2004). A functional polymorphism in the succinate-semialdehyde dehydrogenase (aldehyde dehydrogenase 5 family, member A1) gene is associated with cognitive ability. *Molecular Psychiatry*, *9*, 582–586.
- Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., Beckwith, C. A., Chan, J. A., Hills, A., Davis, M., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with *MYC* in colorectal cancer. *Nature Genetics*, *41*, 882–884.
- Price, A. L., Patterson, N., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909.
- Price, G. R. (1972). Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics*, *36*, 129–140.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575.
- Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.
- Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review*, *9*, 278–291.

- Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology*, *97*, 1756–1774.
- Ratcliff, R. & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.
- Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence*, *36*, 10–17.
- Ratcliff, R. & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333–367.
- Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 127–157.
- Reiss, D., Plomin, R., Hetherington, E. M., Howe, G. W., Rovine, M., Tyron, A., & Hagan, M. S. (1994). The separate worlds of teenage siblings: An introduction to the study of the nonshared environment and adolescent development. In E. M. Hetherington, D. Reiss, & R. Plomin (Eds.), *Separate Social Worlds of Siblings: The Impact of Nonshared Environment on Development* (pp. 63–109). Hillsdale, NJ: Erlbaum.
- Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, *19*, 1736–1753.
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345.
- Roberts, S. & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience* (pp. 611–653). Cambridge, MA: MIT Press.
- Roelfsema, P. R., Lamme, V. A., Spekreijse, H., & Bosch, H. (2002). Figure-ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, *14*, 525–537.
- Rogers, A. R., Iltis, D., & Wooding, S. (2004). Genetic variation at the MC1R locus and the time since loss of human body hair. *Current Anthropology*, *45*, 105–108.

- Sackur, J. & Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. *Cognition*, *111*, 187–211.
- Sajda, P., Philiastides, M. G., Heekeren, H., & Ratcliff, R. (2011). Linking neuronal variability to perceptual decision making via neuroimaging. In M. Ding & D. Glanzman (Eds.), *The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance* (pp. 214–232). New York, NY: Oxford University Press.
- Sanders, A. F. (1990). Issues and trends in the debate on discrete vs. continuous processing of information. *Acta Psychologica*, *74*, 123–167.
- Sanders, A. F. (1998). *Elements of Human Performance: Reaction Processes and Attention in Human Skill*. Mahwah, NJ: Erlbaum.
- Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
- Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414–429.
- Schumaker, R. E. & Lomax, R. G. (2004). *A Beginner's Guide to Structural Equation Modeling* (2nd ed.). Mahwah, NJ: Erlbaum.
- Schweickert, R. & Townsend, J. T. (1989). A trichotomy: Interactions of factors prolonging sequential and concurrent mental processes in stochastic discrete mental (PERT) networks. *Journal of Mathematical Psychology*, *33*, 328–347.
- Sesardic, N. (2005). *Making Sense of Heritability*. Cambridge, UK: Cambridge University Press.
- Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge, MA: Cambridge University Press.
- Shipley, B. (2000). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge, UK: Cambridge University Press.
- Sigman, M. & Dehaene, S. (2005). Parsing a cognitive task: A characterization of the mind's bottleneck. *PLoS Biology*, *3*, e37.
- Sigman, M. & Dehaene, S. (2006). Dynamics of the central bottleneck: Dual-task and task uncertainty. *PLoS Biology*, *4*, e220.

- Sigman, M. & Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *Journal of Neuroscience*, 28, 7585–7598.
- Simons, D. J. & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception*, 28, 1059–1074.
- Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science*, 329, 72–74.
- Song, J.-H. & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, 13, 360–366.
- Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*. New York, NY: Macmillan.
- Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42, 937–948.
- Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search* (2nd ed.). Cambridge, MA: MIT Press.
- Spirtes, P., Richardson, T., Meek, C., Scheines, R., & Glymour, C. (1998). Using path diagrams as a structural equation modelling tool. *Sociological Methods and Research*, 27, 182–225.
- Stacey, S. N., Sulem, P., Zanon, C., Gudjonsson, S. A., Thorleifsson, G., Helgason, A., Jonasdottir, A., Besenbacher, S., Kostic, J. P., Fackenthal, J. D., et al. (2010). Ancestry-shift refinement mapping of the *C6orf97-ESR1* breast cancer susceptibility locus. *PLoS Genetics*, 6, e1001029.
- Sternberg, R. J. (1977). *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities*. Hillsdale, NJ: Erlbaum.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, 30, 276–315.

- Stokowski, R. P., Pant, P. V., Dadd, T., Fereday, A., Hinds, D. A., Jarman, C., Filsell, W., Ginger, R. S., Green, M. R., van der Ouderaa, F. J., et al. (2007). A genomewide association study of skin pigmentation in a South Asian population. *American Journal of Human Genetics*, *81*, 1119–1132.
- Sturm, R. A. (2009). Molecular genetics of human pigmentation diversity. *Human Molecular Genetics*, *18*, R9–R17.
- Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., Martin, N. G., & Montgomery, G. W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color. *American Journal of Human Genetics*, *82*, 424–431.
- Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., et al. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics*, *39*, 1443–1452.
- Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S. A., Palsson, A., Thorleifsson, G., et al. (2008). Two newly identified genetic determinants of pigmentation in Europeans. *Nature Genetics*, *40*, 835–837.
- Telford, C. W. (1931). The refractory phase of voluntary and associative responses. *Journal of Experimental Psychology*, *14*, 1–36.
- Thomson, G. H. (1951). *The Factorial Analysis of Human Ability* (5th ed.). London, UK: University of London Press.
- Thornton, T. L. & Gildea, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review*, *114*, 71–103.
- Turchin, M. C. (2011). Selection for height in Europeans. Presentation at the 2011 CSHL Biology of Genomes Annual Meeting.
- Turkheimer, E. (2008). The gloomy prospect wins: Statistical significance and population stratification in genome wide association studies. *Nature Precedings*.
- Turkheimer, E. & Waldron, M. (2000). Nonshared environment: A theoretical, methodological, and quantitative review. *Psychological Bulletin*, *126*, 78–108.
- Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

- van den Oord, E. J. C. G., Kuo, P. H., Hartmann, A. M., Webb, B. T., Moller, H. J., Hettema, J. M., Giegling, I., Bukszar, J., & Rujescu, D. (2008). Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Archives of General Psychiatry*, *65*, 1062–1071.
- van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*, 381–393.
- van Ravenzwaaij, D. & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*, 463–473.
- Vandekerckhove, J. & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*, 61.
- Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence*, *7*, 53–70.
- Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Genetics Reviews*, *9*, 255–266.
- Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S. E., Hill, W. G., Hottenga, J.-J., Willemsen, G., Boomsma, D. I., Liu, Y.-Z., Deng, H.-W., Montgomery, G. W., & Martin, N. G. (2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics*, *81*, 1104–1110.
- Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory and Cognition*, *32*, 1206–1220.
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671.
- Wagenmakers, E.-J., van der Mass, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin and Review*, *14*, 3–22.
- Wang, K., Dickson, S. P., Stolle, C. A., Krantz, I. D., Goldstein, D. B., & Hakonarson, H. (2010). Interpretation of association signals and identification of causal variants from genome-wide association studies. *The American Journal of Human Genetics*, *86*, 730–742.



- Watson, J. D. & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*, 737–738.
- Welford, A. T. (1980). The single-channel hypothesis. In A. T. Welford (Ed.), *Reaction Times* (pp. 215–252). New York, NY: Academic Press.
- Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661–683.
- Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., & Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, *3*(6), e90.
- Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences USA*, *107*, 5238–5241.
- Wong, K. F. & Wang, X. J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, *26*, 1314–1328.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). Cambridge, MA: MIT Press.
- Wray, N. R., Purcell, S., & Visscher, P. M. (2011). Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biology*, *9*, e1000579.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.
- Wright, S. (1931). Statistical methods in biology. *Journal of the American Statistical Association*, *26*, 155–163.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215.
- Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences USA*, *24*, 253–259.
- Wright, S. (1968). *Evolution and the Genetics of Populations Vol. 1: Genetics and Biometric Foundations*. Chicago, IL: University of Chicago Press.

- Wright, S. (1969). *Evolution and the Genetics of Populations Vol. 2: The Theory of Gene Frequencies*. Chicago, IL: University of Chicago Press.
- Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, *43*, 519–525.
- Yazbek, S. N., Buchner, D. A., Geisinger, J. M., Burrage, L. C., Spiezio, S. H., Zentner, G. E., Hsieh, C. W., Scacheri, P. C., Croniger, C. M., & Nadeau, J. H. (2011). Deep congenic analysis identifies many strong, context-dependent qtls, one of which, *slc35b4*, regulates obesity and glucose homeostasis. *Genome Research*, *21*(7), 1065–1073.
- Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, *329*, 75–77.
- Yuan, K.-H. & Bentler, P. M. (2007). Structural equation modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics Vol. 26: Psychometrics* (pp. 297–358). Amsterdam, The Netherlands: Elsevier.
- Zinkstock, J. R., de Wilde, O., van Amelsvoort, T. A. M. J., Tanck, M. W., Baas, F., & Linszen, D. H. (2007). Association between the DTNBP1 gene and intelligence: A case-control study in young patients with schizophrenia and related disorders and unaffected siblings. *Behavioral and Brain Functions*, *3*, 19.